Chinese Science Bulletin 2004 Vol. 49 No. 12 1311—1313

Reconstruction of the most recent common ancestor sequences of SARS-Cov S gene and detection of adaptive evolution in the spike protein

ZHANG Yuan[1], ZHENG Nan[2,3], HAO Pei[4,5] & ZHONG Yang[6]

1. Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China;
2. College of Life Sciences, Beijing Normal University, Beijing 100875, China;
3. Institute of Viral Disease Control and Prevention, Chinese CDC, Beijing 100052, China;
4. Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China;
5. Shanghai Center for Bioinformation Technology, Shanghai 200235, China;
6. School of Life Sciences, Fudan University, Shanghai 200433, China

Correspondence should be addressed to Zhong Yang (e-mail: yangzhong@fudan.edu.cn)

DOI: 10.1360/04wc0153

The genome organization and expression strategy of severe acute respiratory syndrome coronavirus (SARS-CoV) have been described extensively[1 – 10]. As a structural glycoprotein on the virion surface, the spike protein is responsible for binding to host cellular receptors and for the fusion between the viral envelope and the cellular membrane. It also induces neutralizing antibodies in the host and mediates cellular immunity[11]. Previous studies suggested that amino acid replacements in the spike protein could dramatically alter the pathogenesis and virulence of some coronaviruses[11]. It is therefore reasonable to test the hypothesis that radical amino acid replacements in the spike protein, favored by environmental selective pressure during the process of SARS-CoV interspecific transmission[10], might make this pathogen adapt to a new host. In this study, we investigated a total of 108 complete sequences of the SARS-CoV S gene from GenBank (until March 23, 2004). After omission of those records containing frame-shift mutations or low quality sequences, e.g. ZJ01, and selection of one sequence for identical records, an alignment of 42 sequences was obtained using the program Clustal-X[13]. Then, we reconstructed the most recent common ancestor (MRCA) sequences of the SARS-CoV S gene and detected the adaptive evolution in the spike protein.

The phylogenetic tree of the SARS-CoV S gene was generated using the neighbor-joining (NJ) method[14], implemented in the MEGA2 Package[15] with the Kimura two-parameter distance[16]. In this tree, the viral isolates can be classified into two clusters. One cluster consists of the animal isolates obtained in 2003, and the other cluster comprises all the human isolates during the epidemic from 2002 to 2003. However, the human GD03T0013 isolate was placed between the two clusters, consistent with another recent study[17]. The internal branch linking the two clusters represented the interspecific transmission of SARS-CoV, and the two nodes connected by this branch represented the two clusters' respective MRCAs. It is useful to reconstruct the MRCA sequences and compare them with each other. On the one hand, the comparison between the two inferred MRCA sequences allows us to detect the trends of changes in amino acid properties of the spike protein during the interspecific transmission. On the other hand, it can be synthesized via site-directed mutagenesis technology to express their respective protein products, especially about the affinity with human cellular receptor ACE2[18—20]. Using the codon-based substitution model[21] and the maximum likelihood (ML) method[22] with the assumption of an independent nonsynonymous/synonymous ratio for each branch, we reconstructed an ancestral gene sequence at each node of the phylogenetic tree. The computation was conducted using the CODEML program implemented in the PAML 3.14 package[23]. The overall accuracy of the inferred sequences ranged from 0.95 to 1.00. The comparison between the two MRCA sequences revealed that there are 13 nonsynonymous substitutions but no synonymous substitution (Table 1). Interestingly, the 13 nonsynonymous substitutions were nested in the 28 nonsynonymous substitutions found by comparison among 66 complete sequences in a recent study[17], indicating that our inferred MRCA sequences were reliable and could effectively narrow our focus on nonsynonymous substitutions potentially involved in adaptive evolution. Among the 13 corresponding amino acid replacements, eight were found in the S1 region and five in the S2 region of the spike protein[16—18]. In particular, three amino acid replacements (sites 360, 479 and 487) were located in the receptor-binding region[18,19].

To evaluate the influences of environmental selective pressure on the amino acid properties of the spike protein, both the χ^2 goodness-of-fit (GF) test and the Z test[24] implemented in the program TreeSAAP[25] were employed to detect the trends of changes in 31 biochemical and structural amino acid properties for the 13 nonsynonymous substitutions. In the GF test, according to the change magnitude of a specific property, nonsynonymous substitutions were equally divided into two categories, i.e. conservative (weak changes in this property) and radical (strong changes in this property). Based on the codon composition of gene, the expected substitution frequencies of the two categories were calculated with the assumption of completely random amino acid replacement, i.e.

Table 1 Site differences between the two MRCAs in DNA and amino acid sequences*

DNA sequences			Amino acid sequences		
Site	MRCAa	MRCAh	Site	MRCAa	MRCAh
681	A	C	227	Lys	Asn
782	A	C	261	Lys	Thr
931	G	A	311	Gly	Arg
1079	C	T	360**	Ser	Phe
1437	A	T	479**	Lys	Asn
1460	G	C	487**	Ser	Thr
1819	C	T	607	Pro	Ser
1994	C	T	665	Ser	Leu
2102	T	C	701	Leu	Ser
2227	G	A	743	Ala	Thr
2261	T	C	754	Val	Ala
2680	G	A	894	Ala	Thr
3487	G	A	1163	Glu	Lys

* MRCAa represents the most recent common ancestor of the animal isolates, and MRCAh represents the most recent common ancestor of the human isolates; ** The site is located in the receptor-binding region.

assuming selective neutrality. The expected substitution number of each category was also calculated when the total number of nonsynonymous substitutions was known by sequence comparison. The observed distribution of property change magnitude was then compared with the expected distribution using a χ^2 GF test with the degree of freedom equal to one. A GF score more than 3.84 (α = 0.05) indicated a significant difference between expected and observed distributions and the presence of selective effect. In the *Z* test, the ratio of observed substitution number to all possible substitution number in each category and related standard error were calculated, and then a standard *Z* test was employed to compare the two ratios. The ratio of radical category was expected to be higher than that of conservative category under the assumption of selection effect. A *Z* score more than 1.645 (α = 0.05) suggested the radical change of a specific property favored by natural selection, and *vice versa*. The two tests of the 13 nonsynonymous substitutions revealed selection pressure favoring radical change in four properties: bulkiness, chromatographic index, solvent accessible reduction ratio, and turn tendencies. Our study, however, did not find any selection effect on the other 27 properties. Furthermore, it showed the simultaneous amino acid replacements in three sites: 360 (located in the receptor-binding region), 665 and 701. These sites led to the excess of observed radical substitution number over corresponding expectation under the assumption of selective neutrality, indicative of potentially important roles they played in the adaptive evolution of the spike protein. This finding might be helpful for better understanding of the adaptation mechanism of SARS-CoV and for a basis for the development of anti-SARS drugs and vaccines.

Acknowledgements This work was supported by the National Key Basic Research Special Foundation Project ("973") (Grant No. 2003CB715904).

References

1. Peiris, J., Lai, S., Poon, L. et al., Coronavirus as a possible cause of severe acute respiratory syndrome, Lancet, 2003, 361(9366): 1319—1325.
2. Lu, Y., Chen, Y. H., Spike protein homology between the SARS-associated virus and murine hepatitis virus implies existence of a putative receptor-binding region, Chi. Sci. Bull., 2003, 48(11): 1115—1117.
3. Yu, X. J., Luo, C., Lin, J. C. et al., Putative hAPN receptor binding sites in SARS_CoV spike protein, Acta Pharmaco. Sin., 2003, 24(6): 481—488.
4. 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(20): 1953—1966.
5. Drosten, C., Gunther, S., Preiser, W. et al., Identification of a novel coronavirus in patients with severe acute respiratory syndrome, N. Engl. J. Med., 2003, 348(20): 1967—1976.
6. Kuiken, T., Fouchier, R. A., Schutten, M. et al., Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome, Lancet, 2003, 362(9380): 263—270.
7. 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(5624): 1394—1399.
8. Marra, M. A., Jones, S. J. M., Astell, C. R. et al., The genome sequence of the SARS-associated coronavirus, Science, 2003, 300(5624): 1399—1404.
9. Snijder, E. J., Bredenbeek, P. J., Dobbe, J. C. et al., Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage, J. Mol. Biol., 2003, 331(5): 991—1004.

10. Thiel, V., Ivanov, K. A., Putics, A. et al., Mechanisms and enzymes involved in SARS coronavirus genome expression, J. Gen. Virol., 2003, 84: 2305—2315.
11. Knipe, D. M., Howley, P. M., Fields Virology, Philadelphia, PA: Lippincott Williams & Wilkins Publishers, 2001.
12. Guan, Y., Zheng, B. J., He, Y. Q. et al., Isolation and characterization of viruses related to the SARS coronavirus from animals in Southern China, Science, 2003, 302(5643): 276—278.
13. Thompson, J. D., Higgins, D. G., Gibson, T. J., CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucle. Acid. Res., 1994, 22(22): 4673—4680.
14. Saitou, N., Nei, M., The neighbor-joining method: A new method for reconstructing phylogenetic trees, Mol. Biol. Evol., 1987, 4(4): 406—425.
15. Kumar, S., Tamura, K., Jakobsen, I. B. et al., MEGA2: molecular evolutionary genetics analysis software, Bioinformatics, 2001, 17(12): 1244—1245.
16. Kimura, M., The Neutral Theory of Molecular Evolution, Cambridge (UK): Cambridge University Press, 1983.
17. The Chinese SARS Molecular Epidemiology Consortium, Molecular evolution of the SARS coronavirus during the course of the SARS epidemic in China, Science, 2004, 303(5664): 1666—1669.
18. Li, W., Moore, M. J., Vasilieva, N. et al., Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus, Nature, 2003, 426(6965): 450—454.
19. Wong, S. K., Li, W., Moore, M. J. et al., A 193-amino acid fragment of the SARS coronavirus S protein efficiently binds angiotensin-converting enzyme 2, J. Biol. Chem., 2004, 279(5): 3197—3201.
20. Sui, J., Li, W., Murakami, A. et al., Potent neutralization of severe acute respiratory syndrome (SARS) coronavirus by a human mAb to S1 protein that blocks receptor association, Proc. Natl. Acad. Sci. U S A, 2004, 101(8): 2536—2541.
21. Goldman, N., Yang, A., A codon-based model of nucleotide substitution for protein-coding DNA sequences, Mol. Biol. Evol., 1994, 11(5): 725—736.
22. Yang, Z., Kumar, S., Nei, M., A new method of inference of ancestral nucleotide and amino acid sequences, Genetics, 1995, 141(4): 1641—1650.
23. Yang, Z., PAML: A program package for phylogenetic analysis by maximum likelihood, Comput. Appl. Biosci., 1997, 13: 555—556.
24. McClellan, D. A., McCracken, K. G., Estimating the influence of selection on the variable amino acid sites of the cytochrome B protein functional domains, Mol. Biol. Evol., 2001, 18(6): 917—925.
25. Woolley, S., Johnson, J., Smith, M. J. et al., TreeSAAP: selection on amino acid properties using phylogenetic trees, Bioinformatics, 2003, 19(5): 671—672

(Received April 5, 2004; accepted May 19, 2004)