Chinese Science Bulletin 2003 Vol. 48 No. 12 1175—1178

Phylogeny of SARS-CoV as inferred from complete genome comparison

QI Zhen[1*], HU Yu[2*], LI Wei[1,3], CHEN Yanjun[1], ZHANG Zhihua[1], SUN Shiwei[2], LU Hongchao[2], ZHANG Jingfen[2], BU Dongbo[2], LING Lunjiang[1] & CHEN Runsheng[1,2,3]

1. Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;
2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;
3. Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China

* These authors contributed equally to this work

Correspondence should be addressed to Chen Runsheng (e-mail: crs@sun5.ibp.ac.cn)

Abstract **SARS-CoV, as the pathogeny of severe acute respiratory syndrome (SARS), is a mystery that the origin of the virus is still unknown even a few isolates of the virus were completely sequenced. To explore the genesis of SARS-CoV, the FDOD method previously developed by us was applied to comparing complete genomes from 12 SARS-CoV isolates to those from 12 previously identified coronaviruses and an unrooted phylogenetic tree was constructed. Our results show that all SARS-CoV isolates were clustered into a clique and previously identified coronaviruses formed the other clique. Meanwhile, the three groups of coronaviruses depart from each other clearly in our tree that is consistent with the results of prevenient papers. Differently, from the topology of the phylogenetic tree we found that SARS-CoV is more close to group 1 within genus coronavirus. The topology map also shows that the 12 SARS-CoV isolates may be divided into two groups determined by the association with the SARS-CoV from the Hotel M in Hong Kong that may give some information about the infectious relationship of the SARS.**

Keywords: SARS-associated coronavirus (SARS-CoV), coronavirus, phylogeny, infection relationship, function of degree of disagreement (FDOD).

DOI: 10.1360/03wc0253

SARS-CoV, as the pathogeny of severe acute respiratory syndrome (SARS), seems to be the first coronavirus that is lethal to humans. Coronavirus (family Coronaviridae, genus *Coronavirus*) is an enveloped, single-stranded plus sense RNA virus whose genome has approximately 30 kb size. Whereas coronaviruses may cause severe disease in animals, coronaviruses human strains only cause mild diseases until SARS-CoV was discovered.

To date, SARS-CoV genomes from 12 isolates have been completely sequenced and released[1—4]. Preliminary analysis of SARS-CoV genome indicated that the virus is not phylogenetic closely related to any previously identified coronaviruses. Few obvious clues were given by the genome sequence to answer an important question: what is the origin of SARS-CoV? Based on alignment of amino acid sequences or nucleotide acid sequences, some hypotheses were brought forward to elucidate the origin of SARS-CoV. However, the distant relationship of SARS-CoV to any known virus inferred from very low score of alignment makes these assumptions worthless.

Coronaviruses were classified into three groups according to the serotypes: groups 1 and 2 contain mammalian viruses, while group 3 contains only avian viruses[5,6]. Based on the analysis of phylogeny from predicted proteins of SARS-CoV, Rota et al.[2] claimed that SARS-CoV does not closely resemble any of these three groups and suggest the 4th group for SARS-CoV. Some other authors arrived the same conclusion by analyzing proteins of other isolates[1,3,4].

To resolve the uncertain infectious relationship among different SARS-CoV strains, Ruan et al.[3] compared genomes from 14 SARS-CoV isolates and identified 129 sequence variations among them. Combined with the knowledge of contact source history and geography, common variant sequences were used as genetic signatures to reconstruct a probable lineage map of the SARS-CoV infections. They concluded that the case associated with infections originating in Hotel M in Hong Kong form a group while other isolates form the other one. However, some details are still unclear. Meanwhile, due to the limitation of data, Ruan et al. have to restrict their research to 26140 loci. In addition, some of these 129 mutations might have occurred during *in vitro* expansion and might be sequencing errors rather than true ones[7].

To address such issues, a theoretical method (named FDOD) that we previously developed based on Shannon’s definition of information, entropy and degree of disagreement is used. FDOD calculates species specific complete information set (CIS) from its primary sequence of whole genome, thus circumambulate alignment and avoid any bias that may be associated with particular genomic regions. Primary sequence of a genome is the result of its evolutionary history. The more closely phylogenetic related two species are, the more similar sequences they should have. Hence, CIS can be regarded as a reasonable measure of species distance[8,9].

The software we developed and the supplementary material are available upon request.

1 Materials and methods

To date, genomes from 12 SARS-CoV isolates and 12 previously identified coronaviruses have been completely sequenced. We download these genomes from anonymous ftp server (ftp://130.14.22.5/genbank/genomes). Table 1 gives the related information such as accession number, host, source, group, etc.

Table 1 Information related to 12 SARS-CoV isolates and 12 previously identified coronaviruses

ID	Accession No.	Host	Source	Group	Revision date
cAvian	NC_001451.1	Avian		3	19-NOV-2002
cBovine_1	AF391541.1	Bovine		2	05-FEB-2002
cBovine_2	AF391542.1	Bovine		2	05-FEB-2002
cBovine_3	U00735.2	Bovine		2	23-APR-2003
cBovine_4	AF220295.1	Bovine		2	01-APR-2003
cHuman	AF304460.1	Human		1	11-JUL-2001
cMouse	AF029248.1	Mouse		2	25-JUL-2000
cMurine_1	AF208066.1	Murine		2	11-MAY-2000
cMurine_2	AF201929.1	Murine		2	03-JAN-2002
cMurine_3	AF208067.1	Murine		2	03-JAN-2002
cPig_1	NC_002306.2	Pig		1	28-APR-2003
cPig_2	NC_003436.1	Pig		1	26-APR-2003
SARS_BJ01	AY278488.2	Human	Beijing		01-MAY-2003
SARS_HK_1	AY282752.1	Human	Hong Kong		07-MAY-2003
SARS_HK_2	AY278491.2	Human	Hong Kong		18-APR-2003
SARS_HK_3	AY278554.2	Human	Hong Kong		14-MAY-2003
SARS_SG_1	AY283794.1	Human	Singapore		09-MAY-2003
SARS_SG_2	AY283795.1	Human	Singapore		09-MAY-2003
SARS_SG_3	AY283796.1	Human	Singapore		09-MAY-2003
SARS_SG_4	AY283797.1	Human	Singapore		09-MAY-2003
SARS_SG_5	AY283798.1	Human	Singapore		09-MAY-2003
SARS_TOR2	AY274119.3	Human	Toronto		16-MAY-2003
SARS_TW1	AY291451.1	Human	Taiwan		14-MAY-2003
SARS_Urban	AY278741.1	Human	USA		21-APR-2003

The primary sequences of coronavirus genomes are subjected to FDOD software to calculate the distance matrix based on their discrepancy of CIS[8]. Then, the NEIGHBOR based on neighbor joining algorithm in PHYLIP 3.6 package was used to construct the unrooted tree from distance matrix. To generate multiple data sets for evaluating robustness of the branches of the tree, we adopted the Jackknife algorithm to randomly resample[10]. Finally, the consensus tree is produced using CONSENSE in PHYLIP 3.6.

2 Results and discussion

The unrooted phylogenetic tree was constructed for genomes from 12 SARS-CoV isolates and that from 12 previously identified coronviruses (Fig. 1). It can be split into two parts at the point indicated by the arrow in Fig. 1. All SARS-CoV isolates are located at one side while 12 coronviruses are at the other part. Consistent with the result of Rota et al.[2], the three groups of coronaviruses depart from each other clearly in our tree. The bootstrap value at the divergent point of these three groups is 92% (bootstrap values higher than 70% correspond to a probability higher than 95%[11]).

Our result indicated that SARS-CoV is closer to group 1 of the coronaviruses than to the other two groups (Fig. 2) by the support with the high bootstrap value (bootstrap value is 97% for clad of group 1 and 81% at divergent point of cPig_1 and SARS_TOR2). Differently, Rota et al. regard SARS-CoV as a distinct group within the genus *Coronavirus* based on alignment of amino acid sequences[1—4]. Prevenient paper shows that poorly conserved or variable-length region is not reliable for phylogeny construction based on alignment[12,13]. Since the similarity between SARS-CoV and coronaviruses was very low[2], the new method would be needed to build the phylogenetic tree. The results are based on our new method that circumambulate alignment of sequences and the results are supported moderately by some evidence related to serology. Ksiazek et al.[14] performed immunohistochemical assays with various antibodies reactive with coronaviruses from three groups and with the immune serum specimen from a SARS patient. Their result demonstrated strongly cytoplasmic and membranous staining of infected cells with antibodies related to coronaviruses of group 1 while no staining was identified with any antibodies related to coronaviruses of groups 2 and 3.

Hemagglutinin esterase gene, which presents in all coronaviruses of group 2 and some of group 3, does not exist in SARS-CoV and coronaviruses of group 1[2], that also support our result that SARS-CoV is closer to coronaviruses of group 1. From the result one may assume that the origin of SARS-CoV may be more related to the coronaviruses of group 1 than to those of the other two groups.

Coronaviruses of group 1 cover a wide range of hosts,

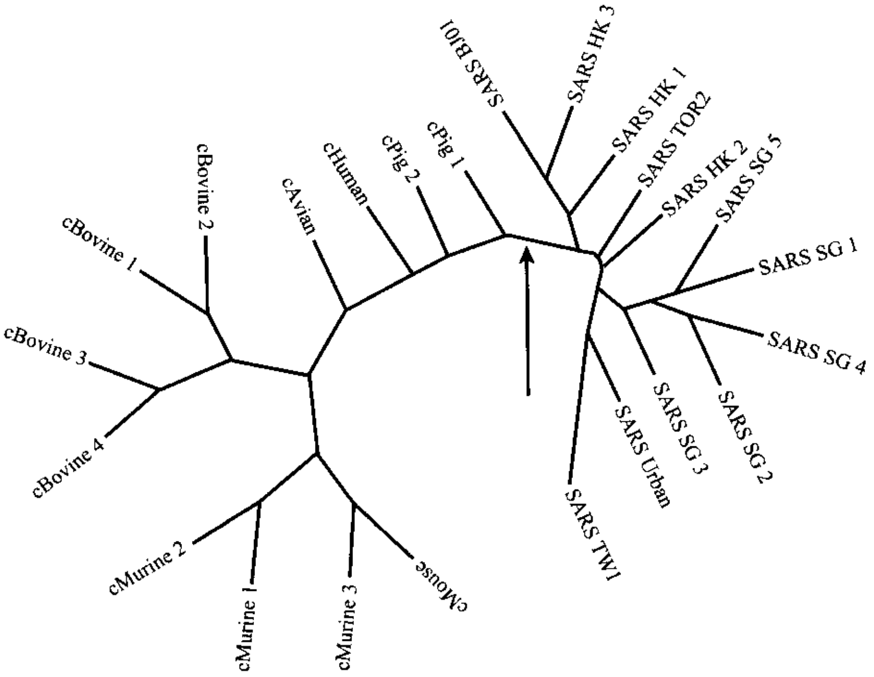

Fig. 1. Phylogenetic tree of 12 SARS-CoV isolates and 12 previously identified coronaviruses. Parameter *K* varies from 3 to 8. Use Jackknife to generate multiple datasets, bootstrap values are the percent of 1000 replicates. Length of clad is proportional to distance based on CIS described above.

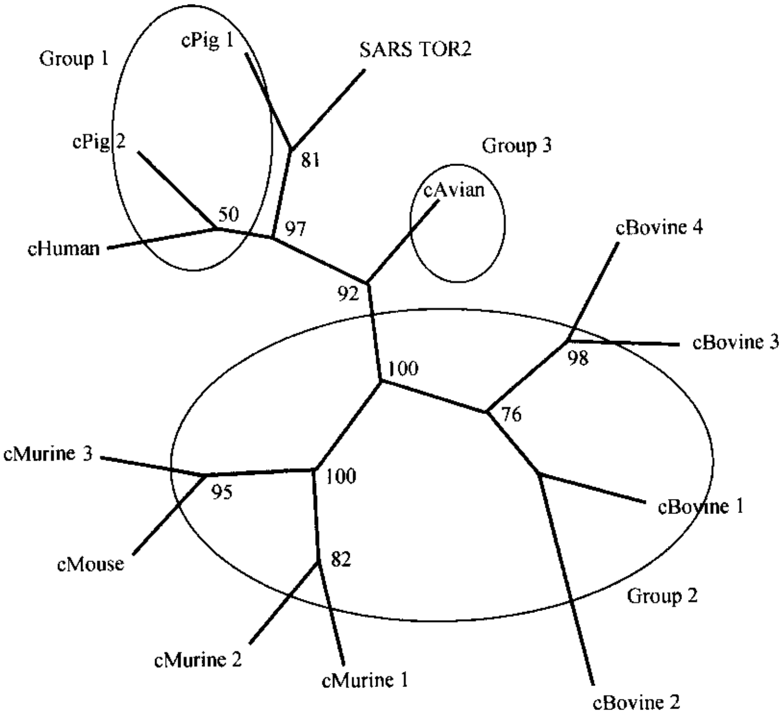

Fig. 2. Phylogenetic tree of SARS_TOR2 isolates of SARS-CoV and 12 previously identified coronaviruses. Parameter *K* varies from 3 to 8. Use Jackknife to generate multiple datasets, bootstrap values are the percent of 1000 replicates. Length of clad is proportional to distance based on CIS described above.

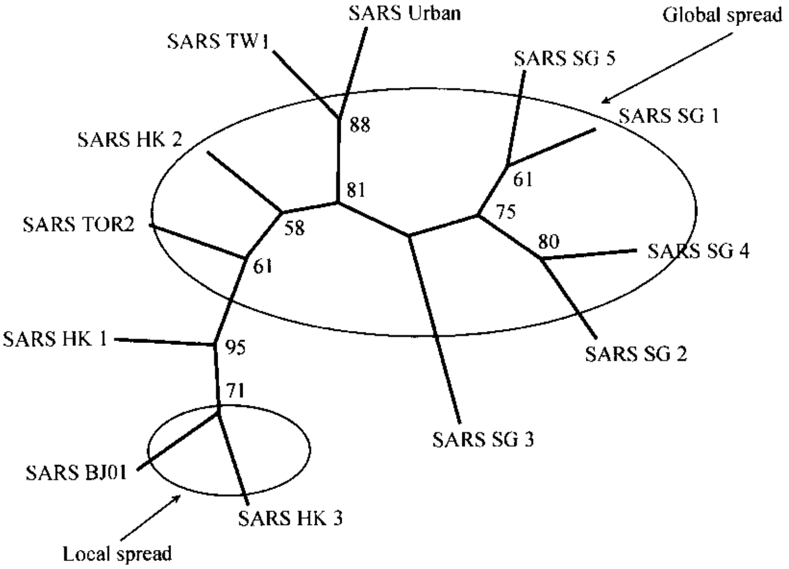

Fig. 3. Phylogenetic relationship of 12 SARS-CoV isolates. Parameter *K* varies from 5 to 8. Use Jackknife to generate multiple datasets, bootstrap values are the percent of 100 replicates. Length of clad is proportional to distance based on CIS described above.

however, we cannot determine which one may be the normal host carrying SARS-CoV. Possibly, SARS-CoV might come from an unknown animal that is not the host of previously identified coronaviruses.

Since the taxon sampling is an important factor influencing the branching pattern of a tree[15], we also construct unrooted trees from different samplings to inspect the robustness of our results and underlying method. Very similar results were acquired that corroborate their robustness of the method (data not shown).

The result shows an infectious relationship map for 12 SARS-CoV isolates which is very similar to that drawn by Ruan et al.[3] (Fig. 3). 12 SARS-CoV isolates form two main groups (designated “Local spread” and “Global spread”) determined by the association with the exposure at Hotel M in Hong Kong. The SARS_SG_3, as a secondary contact case, looks obviously like an ancestral strain among 5 isolates that came from Singapore. However, it is also consistent with the results of Ruan et al. and they attributed it to a potentially back mutational event during the transmission of the virus. SARS_TW1 is placed within the cluster “Global spread” which includes the isolates that are directly or indirectly associated with the infection at Hotel M such as SARS_TOR2, SARS_ HK_2, SARS_Urban, etc. We cannot decide which one of two main groups SARS_HK_1 should be located in, since its contact history is not available.

Acknowledgements This work was supported by the National “863” High-tech Program (Grant No. 2002AA231031), and the Knowledge Innovation Program of the Chinese Academy of Sciences (Grants Nos. KSCX2-2-07 and KJCX1-08).

References

1. Marra, M. A., Jones, S. J. Astell, C. R. et al., The genome sequence of the SARS-associated coronavirus, Science, 2003, 300(5624): 1399—1404.
2. Rota, P. A., Oberste, M. S., Monroe, Stephan S. et al., Characterization of a novel coronavirus associated with severe acute respiratory syndrome, Science, 2003, 300(5624): 1394—1399.
3. Ruan, Y. J., Wei, C. L., Ee, L. A. et al., Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection, Lancet, 2003, 361(9371): 1779—1785.
4. Qin, E. D., Zhu, Q. Y., Wang, J. et al., A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01), Chinese Science Bulletin, 2003, 48(10): 941—948.
5. Enjuanes, L., Brian, D., Cavanagh, D. et al., Coronaviridae, in Virus Taxonomy, Classification and Nomenclature of Viruses (eds. van Regenmortel, M. H. V., Fauquet, C. M., Bishop, D. H. L. et al.), New York: Academic Press, 2000, 835—849.
6. Lai, M. M. C., Holmes, K. V., Coronavmidae: The viruses and their replication, in Fields Virology (eds. Knipe, D. M., Howley, P. M.), 4th ed., New York: Lippincott Williams and Wilkins, 2001, 1163.
7. Bush, R. M., Smith, C. B., Cox, N. J. et al., Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution, Proc. Natl. Acad. Sci. USA, 2000, 7: 6974—6980.
8. Fang, W. W., The characterization of a measure of information discrepancy, Information Sciences, 2000, 125: 207—232.
9. Li, W., Fang, W. W., Ling, L. J. et al., Phylogeny based on whole genome as inferred from complete information set analysis, Journal of Biological Physics, 2002, 28(3): 439—449.
10. Wu, C. F. J., Jackknife, bootstrap and other resampling plans in regression analysis, Annals of Statistics, 1986, 14: 1261—1295.
11. Hillis, D. M., Bull, J. J., An emprical test of bootstrapping as a method for assessing confidence in phylogenetic analysis, Syst. Biol., 1993, 42: 182—192.
12. John, G., Rob, D., Ward, W., Alignment-ambiguous nucleotide sites and the exclusion of systematic data, Mol. Phylogenet. Evol., 1993, 2: 152—157.
13. Ward, W., John, G., Rob, D., Elision: a method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites, Mol. Phylogenet. Evol., 1995, 4: 1—9.
14. Ksiazek, T. G., Erdman, D., Goldsmith, C. et al, A novel coronavirus associated with severe acute respiratory syndrome, N. Engl. J. Med., 2003, 348(20): 1953—1966.
15. Rannala, B., Huelsenbeck, J. P., Yang, Z. et al., Taxon sampling and the accuracy of large phylogenies, Syst. Biol., 1998, 47: 702—710.

(Received May 27, 2003; accepted June 6, 2003)