*Chinese Science Bulletin* 2003 Vol. 48 No. 12 1170—1174

# Molecular phylogeny of coronaviruses including human SARS-CoV

GAO Lei[1,2*], QI Ji[1,2*], WEI Haibin[3,4*], SUN Yigang[3,4] & HAO Bailin[2,4]

1. Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100080, China;
2. T-Life Research Center, Fudan University, Shanghai 200433, China
3. Graduate School, Zhejiang University, Hangzhou 310027, China;
4. Hangzhou Branch, Beijing Genomics Institute, Chinese Academy of Sciences, Hangzhou 310008, China;

Correspondence should be addressed to Hao Baibin (e-mail: hao@itp.ac.cn

* These authors contributed equally to this work.

**Abstract** **Phylogenetic tree of coronaviruses (CoVs) including the human SARS-associated virus is reconstructed from complete genomes by using our newly developed K-string composition approach. The relation of the human SARS-CoV to other coronaviruses, i.e. the rooting of the tree is suggested by choosing an appropriate outgroup. SARS-CoV makes a separate group closer but still distant from G2 (CoVs in mammalian host). The relation between different isolates of the human SARS virus is inferred by first constructing an ultrametric distance matrix from counting sequence variations in the genomes. The resulting tree is consistent with clinic relations between the SARS-CoV isolates. In addition to a larger variety of coronavirus genomes these results provide phylogenetic knowledge based on independent novel methodology as compared to recent phylogenetic studies on SARS-CoV.**

**Keywords: severe acute respiratory syndrome (SARS), coronavirus, molecular phylogeny, composition distance, ultramericity.**

**DOI: 10.1360/03wc0254**

The outbreak of SARS sets an urgent task to reveal the origin of human SARS-CoV, i.e. its relation to other known species of coronavirus, and to trace the genetic variation in the spreading process of SARS. Partial answer to the problem may be obtained from phylogenetic analysis of available genomes. We call a phylogenetic tree of different species of coronavirus including the human SARS-Cov a “CoV Tree” and that of different isolates of SARS-CoV a “SARS Tree”. CoV trees have been constructed by maximal parsimony based on alignment of 405 nt of the CoV polymerase gene ORF 1b[1], and in comparison with predicted amino acid sequences for 6 different proteins[2]. Besides the fact that SARS-CoV makes a separate group with respect to the other three known groups, the precise location of the SARS group remains ambiguous. SARS trees have been built for 5 isolates by aligning complete genomes[3] and for 14 isolates by maximal parsimony based on 16 sequence variations that occurred more than twice[4]. The interrelation of various isolates remains largely uncertain. Moreover, since all SARS genomes sequenced so far are very close to each other, how to construct the SARS Tree requires special consideration. All said calls for a study on more species using an independent methodology. In particular, appropriate choice of an outgroup may provide further indication on where to locate the root of the trees.

## 1 Material and methods

We use 14 complete coronavirus genomes and 17 complete SARS-CoV genomes from GenBank1). Four genomes from Flaviviridae and Togaviridae are used as outgroup. Their abbreviation, accession number and description are given in Table 1.

The CoV Tree is constructed by using our newly developed K-string composition method[5]. This method circumvents alignment of genomic sequences and does not require scoring matrices. It has been successfully applied to prokaryote genomes[5] and chloroplasts[6]. Since this approach yields an unrooted tree, the interrelationship among monophylic groups is examined by adding an outgroup from two distant families of single-strand RNA viruses, Flaviviridae and Togaviridae. Statistical tests of trees built in this way have been discussed in [5] and will not be repeated here.

As regards the SARS Tree the small size of SARS-CoV genome tempts one to align complete genomes for tree construction. However, the high similarity of sequences makes much of the alignment work redundant. In fact, there were only 42 single-letter variations in the first 12 SARS complete genomes (excluding ZJ01, BJ02-4 and GZ01). If one counts the variations among all genome pairs the number varies from 1 to 21. Taking the sequence error rate to be 1 in 100002), there might be 2—3 errors in each genome and 4—6 variations pairwise. Keeping only 16 sequence variations that occurred twice or more as did in [4] is a safe but overcautious approach because there must be single-occurrence variations that are real. If one further excludes the synonymous nucleotide variations these numbers drop from 42/16 to 27/13. Using maximal parsimony means keeping only 16 or 13 variations. Furthermore, the choice of outgroup becomes extremely difficult when all genomes for which we wish to resolve the interrelationship are very close to each other while the candidate outgroup is too distant because an improper outgroup may change the internal branchings in a significant way. In order to make use of all sequence variations at the cost of allowing some sequence errors and to avoid the outgroup problem we propose a new way of tree construction as follows.

1) In the revised paper the 4 partial genomes were replaced by the updated complete ones.

2) The error rate in sequencing BJ01 is estimated to be 0.94 in 10 kb, Private communication from authors of [3].

Table 1 Virus names, abbreviations, NCBI accession numbers and descriptions

| Group | Accession | Abbreviation | Description | |
|---|---|---|---|---|
| G1 | NC_002645.1 | 229E | Human coronavirus 229E, complete genome | |
| G1 | NC_003436.1 | PEDV | Porcine epidemic diarrhea virus strain, complete genome | |
| G1 | NC_002306.2 | TGEV | Transmissible gastroenteritis virus complete genome, genomic RNA | |
| G2 | NC_003045.1 | BCoV | Bovine coronavirus, complete genome | |
| G2 | AF391541.1 | BCoVE | Bovine coronavirus isolate BCoV-ENT, complete genome | |
| G2 | AF391542.1 | BCoVL | Bovine coronavirus isolate BCoV-LUN, complete genome | |
| G2 | U00735.2 | BCoVM | Bovine coronavirus strain Mebus, complete genome | |
| G2 | AF220295.1 | BCoVQ | Bovine coronavirus strain Quebec, complete genome | |
| G2 | NC_001846.1 | MHV | Murine hepatitis virus, complete genome | |
| G2 | AF201929.1 | MHV2 | Murine hepatitis virus strain 2, complete genome | |
| G2 | AF029248.1 | MHVC | Mouse hepatitis virus strain MHV-A59 C12 mutant, complete genome | |
| G2 | AF208067.1 | MHVM | Murine hepatitis virus strain ML-10, complete genome | |
| G2 | AF208066.1 | MHVP | Murine hepatitis virus strain Penn 97-1, complete genome | |
| G3 | NC_001451.1 | IBV | Avian infectious bronchitis virus, complete genome | |
| SARS-CoV | AY278488.2 | BJ01 | SARS coronavirus BJ01, complete genome | |
| SARS-CoV | AY278487.3 | BJ02 | SARS coronavirus BJ02, complete genome | |
| SARS-CoV | AY278490.3 | BJ03 | SARS coronavirus BJ03, complete genome | |
| SARS-CoV | AY279354.2 | BJ04 | SARS coronavirus BJ04, complete genome | |
| SARS-CoV | AY282752.1 | CUHKS | SARS coronavirus CUHK-Su10, complete genome | |
| SARS-CoV | AY278554.2 | CUHKW | SARS coronavirus CUHK-W1, complete genome | |
| SARS-CoV | AY278489.2 | GZ01 | SARS coronavirus GZ01, complete genome | |
| SARS-CoV | AY278491.2 | HKUN | SARS coronavirus HKU-39849, complete genome | |
| SARS-CoV | AY283794.1 | SIN2500 | SARS coronavirus isolate SIN2500 complete genome | |
| SARS-CoV | AY283795.1 | SIN2677 | SARS coronavirus isolate SIN2677 complete genome | |
| SARS-CoV | AY283796.1 | SIN2679 | SARS coronavirus isolate SIN2679 complete genome | |
| SARS-CoV | AY283797.1 | SIN2748 | SARS coronavirus isolate SIN2748 complete genome | |
| SARS-CoV | AY283798.1 | SIN2774 | SARS coronavirus isolate SIN2774 complete genome | |
| SARS-CoV | NC_004718.3 | TOR2 | SARS coronavirus TOR2, complete genome | |
| SARS-CoV | AY291451.1 | TW01 | SARS coronavirus TW1, complete genome | |
| SARS-CoV | AY278741.1 | Urbani | SARS coronavirus Urbani, complete genome | |
| SARS-CoV | AY297028.1 | ZJ01 | SARS coronavirus ZJ01, complete genome | |
| Outgroup | NC_001564 | CellF | Cell fusing agent virus, complete genome | Flaviviridae |
| Outgroup | NC_004102 | HepaCF | Hepatitis C virus, complete genome | |
| Outgroup | NC_001512 | NyongT | O'nyong-nyong virus, complete genome | Togaviridae |
| Outgroup | NC_001544 | RossT | Ross River virus, complete genome | |

If one defines distance between any two species as the branch length to their common ancestor on an additive phylogenetic tree, the distance matrix is ultrametric[9]. Conversely, we may take ultrametricity as a criterion to guide tree construction. A distance matrix derived in some way may not be ultrametric *per se*. However, starting from this matrix one may construct two ultrametric matrices which serve as lower and upper bounds to the original one. In between these two there exist infinitely many ultrametric matrices which may be obtained from the original one by performing various transformations. From these matrices we choose one that is closest to the original one in some well-defined sense as the optimal distance matrix. Starting from this matrix both Unweighted Pair-Group with Arithmetic Mean (UPGMA) or Neighbor-Joining (NJ) (see ref. [7] for these standard methods) would lead to identical trees. The “ultrametrization” has the additional advantage to yield a rooted tree without choosing an outgroup. Actually, choosing an outgroup for the SARS Tree is not a feasible task because all G1 through G3 genomes are too far from the SARS-CoV as it is evident by inspecting the distance matrices. The method of clustering and tree-construction via ultrametrization of distance matrix was sketched in [8]. We implemented the algorithm and applied it to getting the SARS Trees. The method will be described in detail elsewhere and we only present the result in this paper.

## 2 Results and discussion

(i) The CoV Tree. On all 7 CoV trees given in [1] and [2] SARS-CoVs make a separate group besides the 3 known groups. The SARS group is surely distant from G1, but its relation to G2 or G3 varies from tree to tree. In Fig. 1 we present a phylogenetic tree for 20 coronaviruses including 6 SARS-CoVs plus 4 viruses from Flaviviridae and Togaviridae as outgroup. This tree is constructed using composition vectors[5] from the amino acid sequences at string length $K$ = 5.

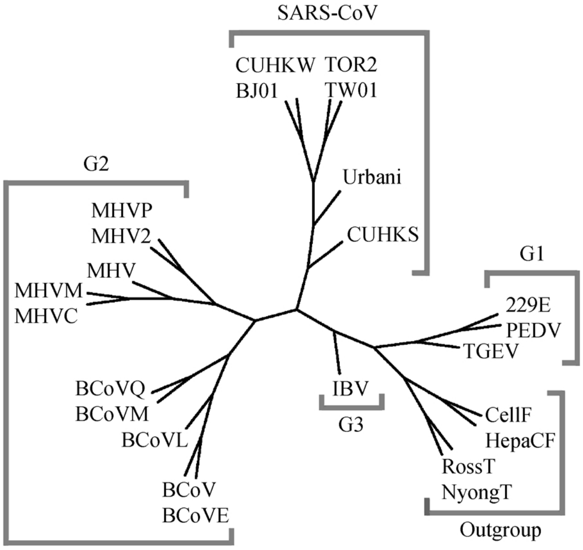

Fig. 1. A phylogenetic tree for 20 coronviruses including 6 SARS-CoVs based on the composition vector method at string length $K$=5. Four viruses from Flaviviridae and Togaviridae are added as outgroup. Note that this is an unrooted tree and the branches are not to scale.

As mentioned above, when monophylic groups on a tree are too distant from each other the intra-group branchings may not be taken seriously as such. One must refer to trees built specially to resolve intra-group relations (see Fig. 2 in Subsection (ii)). The question on SARS origin cannot be answered by phylogenetic study alone as no genomes of close neighbors are present in GenBank for the time being. The only plausible conclusion that may be drawn from all CoV Trees constructed so far is SARS makes a separate group within the Coronavirus genus. The outgroup added to our tree indicates that the SARS group is closer to G2, i.e. to some coronaviruses in mammalian hosts. We mention in passing that the ultrametrization procedure applied to the CoV Tree without using any outgroup also puts the root exactly where the outgroup in Fig. 1 is located.

(ii) The SARS Tree. We first present four distance matrices obtained by counting sequence variations in all available SARS-CoV genomes. The upper right triangle of Table 2 gives pairwise distance by counting all instances of different characters in aligning two sequences (Hamming distance on 4-letter alphabet). There are 137 variations in total. Some nucleotide variations do not change the encoded amino acid if we adopt the Open Reading Frame definitions of the corresponding genome annotation. By excluding these synonymous variations we keep 97 variations shown in the lower left triangle of Table 2. The numbers shown in Table 2 may contain some sequencing errors as well. To be safe one may only keep those sequence variations that occurred twice or more. In this way the numbers 137 and 97 reduce to 18 and 12 without and with synonymous substitutions excluded. These two distance matrices are given in the upper-right and lower-left triangles of Table 3 respectively.

Four SARS Trees built by using the ultrametrization procedure outlined in the Material and methods section are shown in Fig. 2. Fig. 2(a) is based on the 12 sequence variations that occurred at least twice and synonymous substitutions are excluded, i.e. based on the distance matrix given in the lower-left triangle of Table 3. Fig. 2(b) is based on the 18 sequence variations that occurred at least twice but with synonymous substitutions kept. The distance matrix is given in the upper-right triangle of Table 3. Fig. 2(c) is based on all 97 sequence variations including single ones but excluding synonymous substitutions corresponding to the distance matrix given in the lower-left triangle of Table 2. Fig. 2(d) is based on all 137 sequence variations with both single and synonymous ones kept. The distance matrix is given in the upper-right triangle of Table 2.

If the trees built from the 4 distance matrices differ significantly from each other one would not have much to say and more study is required. However, these four trees are topologically consistent in spite of the comparatively large change of the number of variations due to updating of the BJ02-04 and GZ01 genomes from partial to complete. Fig. 2(a) and (b) are based on the most conserved data and turn out to be consistent except for the relocations of CUHKW. They both support the observation[4] that the SARS-CoV spreading process has split into two paths. So does the location of the root. In addition, our method also reveals some finer branches which could not be resolved by using maximal parsimony. We note that these finer branches are consistent with the clinic relations described in [4].

The data used to build trees in Fig. 2(c) and (d) may contain fictitious variations due to sequencing errors, but also make use of real variations that were omitted in Fig. 2(a) and (b). The branchings on these trees are not as reliable as that in Fig. 2(a) and (b). We keep these trees in order to show the improvement reached by excluding single-occurrence variations. The genome sequence of ZJ01 is somehow different from others in that it brings about many more single-sequence variations. However, this does not show off in Fig. 2(a) and (b) when one keeps only variations that occurred twice or more.

We summarize the main findings of this paper. SARS-CoV makes a separate group to the three known groups; its apparent closeness to G2 may be questionable. The origin of SARS-CoV cannot be revealed by phylogenetic study alone at present time as there are too few CoV species represented in GenBank. We must await more

Table 2 Distance matrices based on 137 sequence variations when synonymous substitutions are kept (upper triangle) and on 97 variations when synonymous ones are excluded (lower triangle)

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TW01 | 0 | 6 | 10 | 4 | 3 | 4 | 24 | 2 | 3 | 3 | 4 | 10 | 23 | 23 | 16 | 13 | 50 |
| Urbani | 2 | 0 | 14 | 8 | 7 | 8 | 28 | 6 | 7 | 7 | 8 | 12 | 27 | 27 | 20 | 17 | 54 |
| HKUN | 7 | 9 | 0 | 12 | 11 | 12 | 32 | 10 | 11 | 11 | 12 | 18 | 31 | 31 | 24 | 21 | 58 |
| SIN2677 | 2 | 4 | 9 | 0 | 3 | 4 | 26 | 2 | 5 | 5 | 6 | 12 | 25 | 25 | 18 | 15 | 52 |
| SIN2500 | 2 | 4 | 9 | 2 | 0 | 3 | 25 | 1 | 4 | 4 | 5 | 11 | 24 | 24 | 17 | 14 | 51 |
| SIN2774 | 1 | 3 | 8 | 1 | 1 | 0 | 26 | 2 | 5 | 5 | 6 | 12 | 25 | 25 | 18 | 15 | 52 |
| ZJ01 | 18 | 20 | 25 | 20 | 20 | 19 | 0 | 24 | 25 | 25 | 26 | 32 | 45 | 45 | 38 | 35 | 72 |
| SIN2748 | 1 | 3 | 8 | 1 | 1 | 0 | 19 | 0 | 3 | 3 | 4 | 10 | 23 | 23 | 16 | 13 | 50 |
| TOR2 | 2 | 4 | 9 | 4 | 4 | 3 | 20 | 3 | 0 | 4 | 5 | 11 | 24 | 24 | 17 | 14 | 51 |
| SIN2679 | 1 | 3 | 8 | 3 | 3 | 2 | 19 | 2 | 3 | 0 | 5 | 11 | 24 | 24 | 17 | 14 | 51 |
| CUHKS | 2 | 4 | 9 | 4 | 4 | 3 | 20 | 3 | 4 | 3 | 0 | 10 | 25 | 25 | 18 | 15 | 52 |
| CUHKW | 6 | 8 | 13 | 8 | 8 | 7 | 24 | 7 | 8 | 7 | 8 | 0 | 21 | 21 | 18 | 11 | 46 |
| BJ02 | 16 | 18 | 23 | 18 | 18 | 17 | 34 | 17 | 18 | 17 | 18 | 12 | 0 | 26 | 25 | 16 | 53 |
| BJ03 | 19 | 21 | 26 | 21 | 21 | 20 | 37 | 20 | 21 | 20 | 21 | 15 | 19 | 0 | 25 | 16 | 57 |
| BJ04 | 13 | 15 | 20 | 15 | 15 | 14 | 31 | 14 | 15 | 14 | 15 | 13 | 19 | 22 | 0 | 15 | 54 |
| BJ01 | 9 | 11 | 16 | 11 | 11 | 10 | 27 | 10 | 11 | 10 | 11 | 5 | 9 | 14 | 12 | 0 | 45 |
| GZ01 | 33 | 35 | 40 | 35 | 35 | 34 | 51 | 34 | 35 | 34 | 35 | 27 | 35 | 40 | 38 | 30 | 0 |

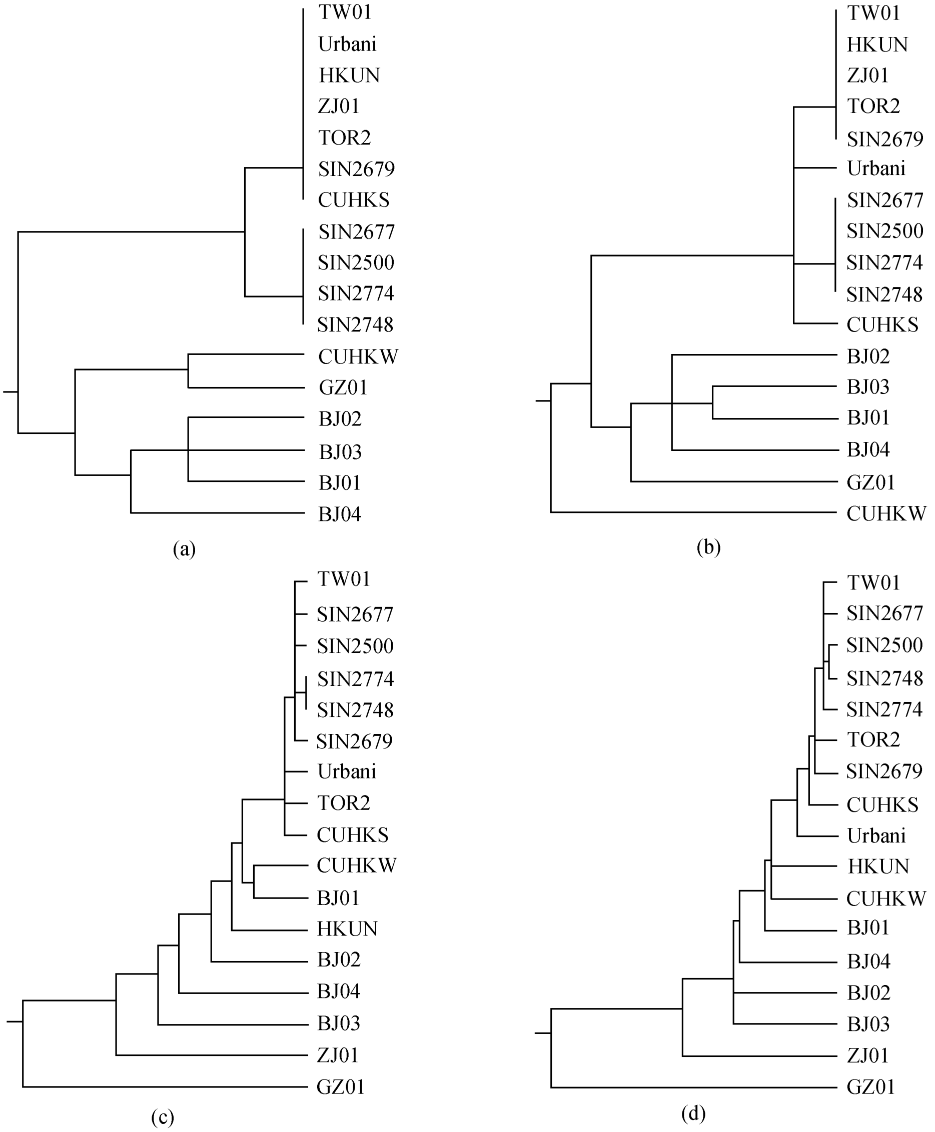

Fig. 2. Ultrametric trees of 17 SARS-CoVs based on sequence variations. (a) All single and synonymous variations excluded (12 remained). (b) Only single variations excluded (18 remained). (c) Single variations kept but synonymous ones excluded (97 remained). (d) All 137 variations are used.

Table 3 Distance matrices based on 18 sequence variations that occurred at least twice (upper right triangle) and on 12 variations when synonymous ones are excluded (lower left triangle)

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TW01 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 8 | 12 | 10 | 6 | 11 | 11 |
| Urbani | 0 | 0 | 1 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 2 | 7 | 13 | 11 | 7 | 12 | 12 |
| HKUN | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 8 | 12 | 10 | 6 | 11 | 11 |
| SIN2677 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 2 | 9 | 13 | 11 | 7 | 12 | 12 |
| SIN2500 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 2 | 9 | 13 | 11 | 7 | 12 | 12 |
| SIN2774 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 2 | 9 | 13 | 11 | 7 | 12 | 12 |
| ZJ01 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 8 | 12 | 10 | 6 | 11 | 11 |
| SIN2748 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 2 | 9 | 13 | 11 | 7 | 12 | 12 |
| TOR2 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 8 | 12 | 10 | 6 | 11 | 11 |
| SIN2679 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 8 | 12 | 10 | 6 | 11 | 11 |
| CUHKS | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 7 | 13 | 11 | 7 | 12 | 12 |
| CUHKW | 6 | 6 | 6 | 7 | 7 | 7 | 6 | 7 | 6 | 6 | 6 | 0 | 10 | 8 | 8 | 9 | 7 |
| BJ02 | 10 | 10 | 10 | 11 | 11 | 11 | 10 | 11 | 10 | 10 | 10 | 6 | 0 | 4 | 6 | 5 | 5 |
| BJ03 | 8 | 8 | 8 | 9 | 9 | 9 | 8 | 9 | 8 | 8 | 8 | 4 | 2 | 0 | 4 | 3 | 7 |
| BJ04 | 5 | 5 | 5 | 6 | 6 | 6 | 5 | 6 | 5 | 5 | 5 | 5 | 5 | 3 | 0 | 5 | 7 |
| BJ01 | 8 | 8 | 8 | 9 | 9 | 9 | 8 | 9 | 8 | 8 | 8 | 4 | 2 | 2 | 3 | 0 | 6 |
| GZ01 | 8 | 8 | 8 | 9 | 9 | 9 | 8 | 9 | 8 | 8 | 8 | 2 | 4 | 4 | 5 | 4 | 0 |

CoV genomes, probably from other mammalians, to be sequenced. A "clinic tree" of SARS spreading like the clinic relation described in [4] does not necessarily imply a phylogenetic tree at molecular level. However, the fact that the SARS Trees (Fig. 2) are consistent with each other and with the clinic relations described in [4] is a manifestation of high mutation rate of SARS-CoV.

**Acknowledgements** The authors thank Li Wei, Beijing Genomics Institute, for providing error estimate of the BJ01 genome; Luo Jingchu, Peking University, for discussion and checking with data on CBI's Anti-SARS website (http://antisars.cbi.pku.edu.cn:5555/index.jsp); Edison T. Liu, Genome Institute of Singapore, for clarifying single- sequence variations; Chu Ka Hou, Chinese University of Hong Kong, for suggesting species for outgroup on the CoV tree. This work was partly supported by the Special Funds for Major State Basic Research Projects, the National Natural Science Foundation of China (Grant No. 30170232), the Innovation Project of the Chinese Academy of Sciences, and by a grant from Shanghai Municipality via Fudan University.

## References

1. Ksiazek, T. G., Erdman, D., Goldsmith, C., et al. A novel coronavirus associated with Severe Acute Respiratory Syndrome, N. Engl. J. Med., 2003, 348: 1953—1966.
2. 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.
3. Qin, E. D., Zhu, Q. Y., Yu, M. et al., A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01), Chinese Science Bulletin, 2003, 48(10): 941—948.
4. Ruan, Y. J., Wei, C. L., Ee, L. A. et al., Comparative full-length genome sequence analysis of 14 SARS coronovirus isolates and common mutations associated with putative origins of infection, The Lancet, 2003, 361(9371): 1779—1785.
5. Qi, J., Wang, B., Hao, B. L., Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach, J. Mol. Evol., 2003 (in print).
6. Chu, K. H., Qi, J., Yu, Z. G. et al., Origin and phylogeny of chloroplasts: a simple correlation analysis of complete genomes, Mol. Biol. Evol. (under revision).
7. Nei, M., Kumar, S., Molecular Evolution and Phylogenetics, New York: Oxford University Press, 2000, 87—103.
8. Rammal, R., Toulouse, G., Virasoro, M., Ultrametricity for physicists, Rev. Mod. Phys., 1986, 58: 765—788.

(Received May 27, 2003; accepted June 12, 2003)