JOURNAL OF VIROLOGY, Feb. 2004, p. 1602–1603 Vol. 78, No. 3
0022-538X/04/$08.00+0 DOI: 10.1128/JVI.78.3.1602–1603.2004

Letter to the Editor

Severe Acute Respiratory Syndrome Coronavirus Sequence Characteristics and Evolutionary Rate Estimate from Maximum Likelihood Analysis

In November 2002, a previously unknown severe acute respiratory syndrome (SARS) was observed in patients of the Guangdong Province, China (7). In March 2003, a new coronavirus (SARS-CoV) was associated with the SARS outbreak (2), and several full-genome sequences of SARS-CoV were obtained and compared (4). The family *Coronaviridae* comprises large, single, plus-stranded RNA viruses isolated from several species and previously known to cause common colds and diarrheal illnesses in humans (3). The emergence of such a novel, highly virulent pathogen warrants rapid investigation of its etiology and evolution to effectively control its impact on human health. In particular, estimating the rate of evolution of SARS-CoV would give an indication of how quickly the virus can potentially increase its genetic variability, which in turn has important implications for disease progression and drug and vaccine development.

Phylogenetic analysis has proven successful for the investigation and prediction of the evolution of viruses such as influenza virus (1). Initial inspection of SARS-CoV sequences revealed a high degree of homogeneity, which might indicate an RNA virus that evolves unusually slowly. To investigate further, we carried out a full-genome alignment of the available SARS-CoV strains recently analyzed by Ruan et al. (4) by use of the CLUSTAL algorithm (6). The alignment was carefully edited by hand to maximize the number of identities, and the site positions containing gaps were removed. The resulting alignment (available from the authors upon request) is 21,333 nucleotides long; 63 sites have at least one sequence with a different nucleotide, and only 10 sites are phylogenetically informative, i.e., they are useful to discriminate among different tree topologies, according to the unweighted parsimony criterion. Subalignments were generated for all of the known coding regions, most of which were identical among the different isolates. We analyzed open reading frame (ORF) 1ab (4), which appears to be the most variable. Maximum likelihood (ML) methods were employed for the analyses because they allow for the testing of different phylogenetic hypotheses by calculating the probability of a given model of evolution generating the observed data and by comparing the probabilities of nested models by the likelihood ratio test (5). In addition, because only 10 sequences were retained after excluding the identical ones, it was possible to search for the optimal ML tree through an exhaustive or branch-and-bound search (5).

Table 1 shows the average base composition and the ML estimates of parameters describing the mode of evolution of SARS-CoV in ORF 1ab. The α parameter of the Γ distribution is extremely low (0.008), implying an extensive heterogeneity in the rate at which different nucleotide sites mutate along the genome. Moreover, the ML estimator implies that about 90% of the constant sites in the sequences are indeed invariable, i.e., they never change, possibly because of strong purifying selection. The variable sites, on the other hand, accumulate mutations very quickly. However, a note of caution is necessary because such a result may also be due to the small number of sequences available for analysis and the very short observation period. Table 1 also shows that the hypothesis of a molecular clock cannot be rejected, although the *P* value is very close to 0.05; i.e., SARS-CoV isolates appear to be evolving at a constant evolutionary rate, which can be estimated from the ML tree with clock-like branch lengths shown in Fig. 1. The branch lengths in the tree are proportional to the number of mutations accumulated by each viral lineage during evolution from the cenancestor, the most recent common ancestor. Assuming that the SARS-CoV cenancestor entered the human population 4 to 8 months ago (7), the evolutionary rate of the virus is of the order of 4×10^{-4} nucleotide changes per site per year (95% confidence interval [CI], 2.0×10^{-4} to 6×10^{-4}) along the entire ORF 1ab. When only the variable sites are considered, the estimated rate is noticeably higher: 3.5×10^{-3} changes per site per year (95% CI, 2.6×10^{-3} to 4.4×10^{-3}). This is the usual range for an RNA virus. Therefore, on average, eight point mutations are expected for the entire ORF 1ab region at each replication round. However, we cannot exclude the possibility that the sequence variability in the data sets is also affected by the passage of the virus in Vero cell culture before sequencing (4). Figure 1 also shows that the root of the tree, inferred by ML, is between the strains isolated from Hong Kong and Beijing, which are known to be epidemiologically linked to the strains isolated from patients in Guangdong Province and all the others (4). Epidemiological data also indicate that the index patient traveled from Guangdong to Hotel M in Hong Kong, where he transmitted the virus to several individuals who successively traveled to Singapore, Canada, and Vietnam (4). The tree shows, indeed, that the Singapore isolate and the isolates from Beijing belong to different, statistically supported clusters. However, because of the low phylogenetic signal, further classification of SARS-CoV isolates is not possible by phylogenetic analysis. All analyses confirm that SARS-

TABLE 1. Maximum likelihood estimators of nucleotide substitution model parameters for the SARS virus in ORF 1ab polyprotein[a]

Tree	ML estimate[b]			Inferred evolutionary rate	
	Transition/transversion rate ratio	α	Pinv	All sites (10^{-4})	Variable sites (10^{-3})
ML[c]	1.68	0.008	0.910		
MLK[d]	1.88	0.0024	0.885	4.0 ± 2.0	3.5 ± 0.9

[a] Percent average base composition was as follows: A, 28.4; C, 19.5; T, 21.3; G, 30.8.

[b] The best-fitting nucleotide substitution model (HKY85+Γ+I) was selected with a hierarchical likelihood ratio test procedure by using a suboptimal tree (5). The model assumes unequal transition and transversion substitution rates, different categories of sites along the genome changing at different rates (described by the α parameter of a Γ distribution of rates), and a class of invariable sites (described by the Pinv parameter).

[c] Estimates are based on the maximum likelihood tree.

[d] The likelihood for each possible rooted tree was obtained, and the best tree, according to the Shimodaira-Hasegawa test, was selected as the maximum likelihood clock tree. The clock hypothesis could not be rejected by the likelihood ratio test ($P = 0.052$).

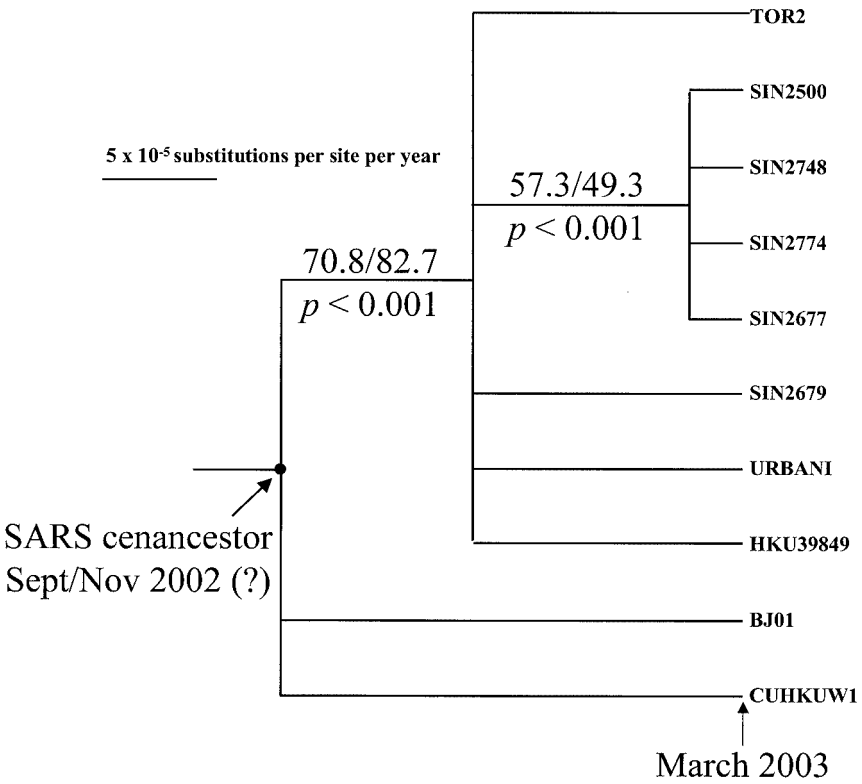

FIG. 1. Optimal ML tree of SARS-CoV ORF 1ab nucleotide sequences. Branch lengths are drawn proportional to the number of nucleotide changes per site and were estimated via ML enforcing a molecular clock and employing the HKY85+Γ+I nucleotide substitution model (Table 1). The numbers on the branches represent the percentages of bootstrap-jackknife support (1,000 replicates) for the subtending clade. The *P* value for the zero-branch-length test (7) is also given.

CoV is not closely related to any known coronavirus (4), although it is assumed that the source must be one or more unidentified animal reservoirs in Asia.

In conclusion, the low sequence variability of SARS-CoV isolates is probably the consequence of its recent emergence in humans, but much greater viral heterogeneity with unpredictable consequences may be expected if the epidemic is not controlled. A rigorous phylogenetic approach might be an important tool to monitor the future evolution of the virus.

REFERENCES

1. **Bush, R. M., C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch.** 1999. Predicting the evolution of human influenza A. Science **286:**1921–1925.
2. **Drosten, C., S. Gunthe, W. Preiser, et al.** 2003. Identification of a novel coronavirus associated with severe acute respiratory syndrome. N. Engl. J. Med. **348:**1967–1976. [Online.]
3. **Holmes, K. V., and M. M. C. Lai.** 1996. *Coronaviridae*: the viruses and their replication, p. 1075–1103. *In* B. N. Fields, D. M. Knipe, and P. M. Hawley (ed.), Fields virology, 3rd ed. Lippincott-Raven Publishers, Philadelphia, Pa.
4. **Ruan, Y., C. L. Wei, A. L. Ee, V. B. Vega, et al.** 2003. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. Lancet **361:**1756–1757.
5. **Swofford, D., and J. Sullivan.** 2003. Phylogenetic inference based on parsimony and other methods with PAUP*. *In* M. Salemi and A.-M. Vandamme (ed.), The phylogenetic handbook—a practical approach to DNA and protein phylogeny. Cambridge University Press, New York, N.Y.
6. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
7. **World Health Organization.** Cumulative of reported probable cases of severe acute respiratory syndrome (SARS). [Online.] http://www.who.int/csr/sarscountry/2003_04_24/en/.

Marco Salemi
Walter M. Fitch
Department of Ecology and Evolutionary Biology
University of California, Irvine
Irvine, California

Massimo Ciccozzi*
Maria Jose Ruiz-Alvarez
Giovanni Rezza
Department of Infectious Parasitic and Immune-Mediated Disease
Istituto Superiore di Sanitá
Rome, Italy

Martha J. Lewis
Department of Internal Medicine
University of California, Los Angeles
Los Angeles, California

*Phone: 0039 0649902337
Fax: 0039 0649387210
E-mail: ciccozzi@iss.it