The Journal of Microbiology, February 2006, p.83-91
Vol. 44, No. 1

Close Relationship Between SARS-Coronavirus and Group 2 Coronavirus

Ok Ju Kim[1], Dong Hun Lee[1,2] and Chan Hee Lee[1,2,*]

Division of Life Sciences[1] and Research Institute for Biotechnology[2], Chungbuk National University, Cheongju, Chungbuk 361-763, Republic of Korea

(Received October 26, 2005 / Accepted January 1, 2006)

The sudden appearance and potential lethality of severe acute respiratory syndrome (SARS)-associated coronavirus (SARS-CoV) in humans has resulted in a focusing of new attention on the determination of both its origins and evolution. The relationship existing between SARS-CoV and other groups of coronaviruses was determined via analyses of phylogenetic trees and comparative genomic analyses of the coronavirus genes: polymerase (Orf1ab), spike (S), envelope (E), membrane (M) and nucleocapsid (N). Although the coronaviruses are traditionally classed into 3 groups, with SARS-CoV forming a 4th group, the phylogenetic position and origins of SARS-CoV remain a matter of some controversy. Thus, we conducted extensive phylogenetic analyses of the genes common to all coronavirus groups, using the Neighbor-joining, Maximum-likelihood, and Bayesian methods. Our data evidenced largely identical topology for all of the obtained phylogenetic trees, thus supporting the hypothesis that the relationship existing between SARS-CoV and group 2 coronavirus is a monophyletic one. Additional comparative genomic studies, including sequence similarity and protein secondary structure analyses, suggested that SARS-CoV may bear a closer relationship with group 2 than with the other coronavirus groups. Although our data strongly suggest that group 2 coronaviruses are most closely related with SARS-CoV, further and more detailed analyses may provide us with an increased amount of information regarding the origins and evolution of the coronaviruses, most notably SARS-CoV.

Keywords: SARS, coronavirus, phylogeny

The sudden outbreak of a previously unknown disease, termed "Severe Acute Respiratory Syndrome (SARS)", during the first half of 2003, prompted a massive search for the causative agent of this disease. The results of extensive study indicated that the causative agent of SARS was a member of the *Coronavirus* genus (Drosten *et al.*, 2003; Ksiazek *et al.*, 2003; Peris *et al.*, 2003), hence referred to as the SARS-associated coronavirus (SARS-CoV). Molecular biological analyses of SARS-CoV identified 14 open reading frames (ORFs) (Marra *et al.*, 2003; Thiel *et al.*, 2003). A proportion of the ORFs discovered in the SARS-CoV genome are common to all coronavirus families, although some are unique to SARS-CoV (Marra *et al.*, 2003; Rota *et al.*, 2003).

The search for the origins of SARS-CoV indicated that SARS-CoV evolved from an animal coronavirus, and this hypothesis was supported by the isolation of SARS-CoV-like coronaviruses from Himalayan palm civets, and from a species of "raccoon dog" sold in a wild animal market in Guangdong, China, the region in which the index case of SARS is presumed to have emerged in humans (Guan *et al.*, 2003). Recently, SARS-like coronaviruses have also been isolated from bats (Lau *et al.*, 2005; Poon *et al.*, 2005). Since the initial full genomic sequencing of SARS-CoV strain Tor2, which was achieved by Marra *et al.* in 2003, the complete or partial genome of SARS-CoV has been sequenced at several sites around the world. As of October 7, 2005, the NCBI GenBank database harbored 126 complete genome sequences for SARS-CoV, including those of both human and animal origin. Using phylogenetic approaches with full or partial sequences of SARS-CoV, as well as other coronaviruses, earlier studies succeeded in separating SARS-CoV from other coronaviruses, and it is currently generally accepted that SARS-CoV constitutes a 4th group in the *Coronavirus* genus (Ksiazek *et al.*, 2003; Marra *et al.*, 2003; Rota *et al.*, 2003). Thus, the genus *Coronavirus* can be classed into 4 groups; mammalian groups 1 and 2, avian group 3, and SARS-CoV group 4. Further attempts to delimit the location of SARS-CoV within phylogenetic trees in accordance with sequence similarity, however, have failed to gen-

* To whom correspondence should be addressed.
(Tel) 82-43-261-2304; (Fax) 82-43-273-2451
(E-mail) chlee@chungbuk.ac.kr

erate a definitive consensus. Although many studies have identified the group 2 coronaviruses as the closest relatives of SARS-CoV (Eickmann *et al.*, 2003; Snijder *et al.*, 2003; Gibbs *et al.*, 2004; Zhu and Chen, 2004), these results have been disputed by other study groups. For example, some reports have asserted that SARS-CoV is closely related with the group 3 or group 1 coronaviruses (Marra *et al.*, 2003; Yap *et al.*, 2003). Furthermore, some arguments have militated for a mosaic origin of SARS-CoV (Rest and Mindel, 2003; Stavrinides and Guttman, 2004). These discrepancies with regard to the origin and relative proximity of SARS-CoV to other coronavirus groups may be attributable to regional variations, or to the methodologies utilized in the individual analyses.

In this study, we have conducted analyses of the full sequences of representative strains of the four groups in the *Coronavirus* genus, via a variety of methods. Phylogenetic trees constructed by the three most frequently employed methods, similarity analysis and predicted secondary structures, were generally suggestive of the notion that the group 2 corona viruses represent the closest relatives to SARS-CoV.

Materials and Methods

Sequences

Nucleotide and amino acid sequences of representa-

Table 1. Sequences analyzed in this study

NCBI GenBank Acc. No.	Group	Name of virus strain	Size (nt)
NC_002306	1	TGEV Purdue	28,586
NC_003436	1	PEDV CV777	28,033
AY518894	1	HCoV Pneumonia	27,555
NC_005831	1	HCoV NL63	27,553
NC_002645	1	HCoV 229E	27,317
AB088223	1	FIPV (S gene)	4,389
AB086903	1	FIPV (M, N genes)	1,938
AY342160	1	CCoV (S gene)	9,203
AF207551	2	RtCoV (S, E, M, N genes)	9,859
AF481863	2	PHEV (S, E, M, N genes)	8,123
NC_001846	2	MHV A59	31,357
L07748	2	HCoV 4408 (S gene)	4,133
AY316299	2	HCoV 4408 (E, M, N genes)	3,037
NC_005147	2	HCoV OC43	30,738
NC_003045	2	BcoV ENT	31,028
AY342357	3	TCoV (S gene)	3,711
AF111995	3	TCoV (N gene)	1,732
AY319651	3	IBV-BJ	27,733
NC_001451	3	IBV-Beaudette	27,608
NC_004718	4	SARS-CoV Tor2	29,751
AY304486	4	SARS-CoV SZ3	29,741
AY304488	4	SARS-CoV SZ16	29,731
AY291315	4	SARS-CoV Frankfurt 1	29,727

Abbreviations : nt, nucleotide; TGEV, Transmissible gastroenteritis virus; PEDV, Porcine epidemic diarrhea virus; HCoV, Human coronavirus; FIPV, Feline infectious peritonitis virus; CCoV, Canine coronavirus; MHV, Murine hepatitis virus; BCoV, Bovine coronavirus; PHEV, Porcine hemagglutinating encephalomyelitis virus; HECoV, Human enteric coronavirus; RtCoV, Rat sialodacryoadenitis coronavirus; IBV, Avian infectious bronchitis virus: SARS-CoV. SARS coronavirus

tive viruses from each of the four coronavirus groups were acquired from the NCBI GenBank. A total of 16 sequences, comprising the entirety of the *Coronavirus* genus, were obtained (Table 1). As longer sequences tend to generate more information than can be gleaned from shorter sequences, the sequences for S, E, M and N were concatenated into structural (ST) genes, and the Orf1a and Orf1b sequences were combined into non-structural (NS) genes. The ST and NS gene sequences were then further concatenated, resulting in the formation of the full sequences. Thus, the term "full sequence" as used in this article, does not refer to the complete genomic sequence, but rather to the following concatenated coding sequence; 5'-Orf1a-Orf1b-S-E-M-N-3'.

Phylogenetic tree construction

The sequences were aligned with the ClustalX program (ver. 1.81), at default parameters. The output data were utilized in the generation of phylogenetic trees, via three methods. Neighbor-joining (NJ) trees were constructed with PHYLIP version 3.6a software (J. Felsenstein, Phylogeny Inference Package, Department of Genetics, University of Washington, Seattle) using the SEQBOOT, DNADIST, PROTDIST, NEIGHBOR, and CONSENSE programs. Protein distance calculations were predicated on the Jones-Taylor-Thornton protein weight matrix, with 1,000 bootstrap (BS) replicates. All other variables were set to default values. Maximum likelihood (ML) trees were generated using the DNAML and PROML programs in PHYLIP. Bayesian (BA) trees were constructed using MRBAYES version 3.0 (http://morphbank.ebc.uu.se/mrbayes), using a Jones-Taylor-Thornton protein weight matrix with 500,000 generations, and a burn-in of 100. Consensus trees were generated via majority rule, and visualized with TREEVIEW (ver 1.6.6).

Similarity analysis

The nucleotide and amino acid sequences of the coronavirus genes obtained from NCBI GenBank (Table 1) were multiple-aligned using ClustalX. The aligned sequences were then compared using the SeqAid software (ver 0.91), allowing for the acquisition of similarity values between sequences belonging to two different coronavirus groups. For example, if there are 4 sequences in SARS-CoV and 7 sequences in a group 1 coronavirus, 28 (4×7) similarity values would be obtained. Then, mean and standard error of the mean values for each pair of comparisons were calculated, and the results were plotted on a bar graph, using SigmaPlot (ver. 8.0). Statistical analyses were conducted using SPSS (ver. 1.0) in order to determine the statistical significance of the data.

Analysis of the protein secondary structures

The secondary structures of the 2 structural (Orf1a, Orf1b) and 4 nonstructural (S, E, M, N) coronavirus proteins were predicted using the HNN program of the ExPASy (Expert Protein Analysis System, http://us.expasy.org/tools/) software package. The predicted secondary structures included alpha helices, extended strands, and random coils. The relative proportion of the 3 secondary structures of SARS-CoV proteins were compared with those of other coronavirus groups, and plotted on a scattergram, using SigmaPlot (ver 8.0). Regression analysis was conducted for each of the scattergrams, and each scattergram contained a regression line with an equation, 95% confidence interval lines, and r^2 (regression coefficient).

Results

Construction and analysis of the phylogenetic trees

The nucleotide sequences of representative coronaviruses from each of the previously established three groups, and of the newly-discovered 4th group (SARS-CoVs) were obtained from NCBI GenBank (Table 1). The acquired sequences included 2 nonstructural genes (Orf1a, Orf1b) and 4 structural genes (S, E, M, N), which are common to all coronaviruses. The amino acid sequences were deduced via the translation of the nucleotide sequences. Then, two concatenated sequences were generated from Orf1a+Orf1b (nonstructural, NS), and S+E+M+N (structural, ST). Finally, the NS and ST sequences were concatenated together, resulting in the formation of full sequences. The concatenated amino acid sequences were generated via the combination of the amino acid sequences of each of the gene, rather than by the translation of the corresponding concatenated nucleotide sequences. A total of 9 natural and concatenated gene sequences were subjected to phylogenetic tree construction. The trees were constructed using the most popular methods, namely the Neighbor-joining (NJ), Maximum-likelihood (ML), and Bayesian (BA) techniques. Thus, a total of 54 phylogenetic trees, 27 from the nucleotide sequences and 27 from the amino acid sequences, were generated and subjected to analysis.

All 54 of the phylogenetic trees, without exception, demonstrated 4 clearly distinctive clusters, corresponding to the 4 coronavirus groups: mammalian group 1 coronavirus (G1-CoV), mammalian group 2 coronavirus (G2-CoV), avian group 3 coronavirus (G3-CoV), and group 4, or SARS-CoV. Due to space limitations, only trees constructed on the basis of full sequences are shown in Fig. 1. The majority of the topologies of the 54 trees were strongly supported by the results of bootstrap analyses conducted on the NJ trees. Most of the branches separating the 4 groups also scored more

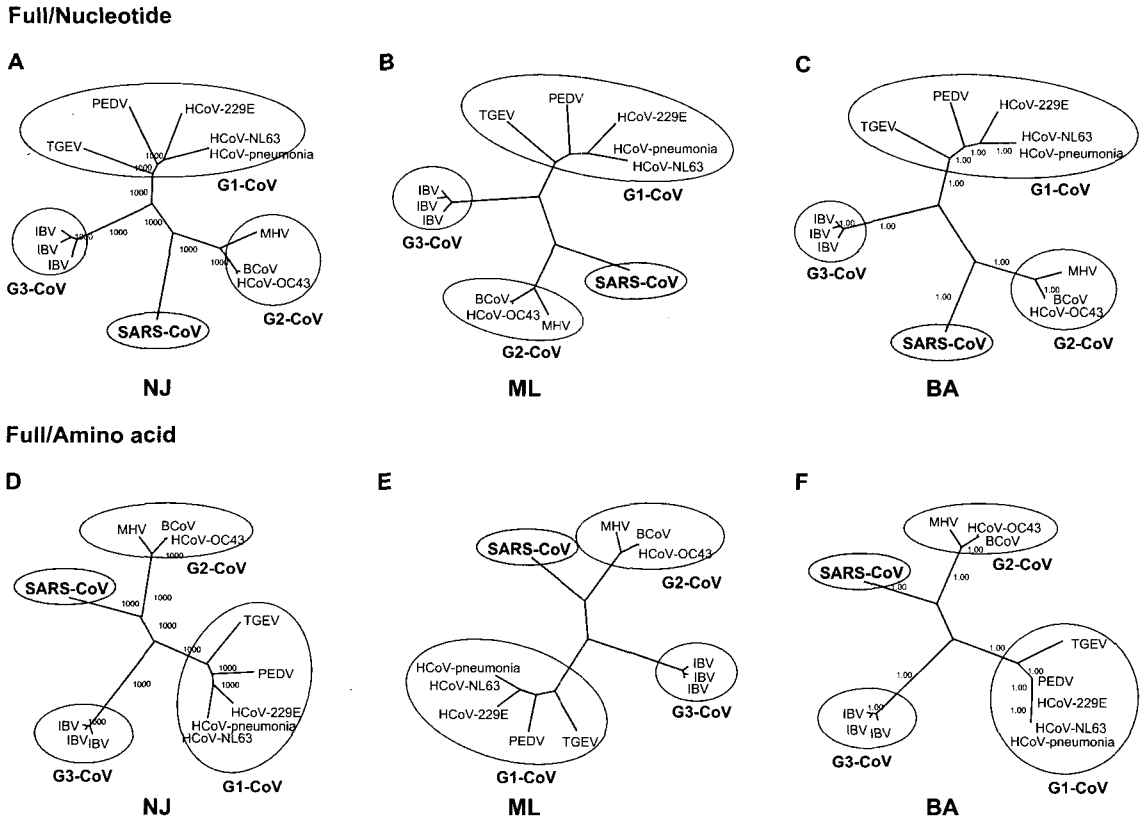

Fig. 1. Unrooted trees of coronaviruses, including SARS-CoV. Full nucleotide or amino acid sequences obtained (as in Table 1) were aligned, and the unrooted trees were generated via Neighbor-Joining (NJ: A, D), Maximum-Likelihood (ML: B, E) or Bayesian (BA: C, F) methods. Bootstrap values for the major branches are shown in the NJ and BA trees.

Table 2. Coronavirus group placed as the nearest neighbor of SARS-CoV in phylogenetic tree topology

Sequence	Tree type	Orf									
		Full	NS[1]	ST[2]	Orf1a	Orf1b	S	E	M	N	E/M
Nucleotide	NJ	G2	G2	G2	G2	G2	G2	G1	G3	G2	G2
	ML	G2	G2	G2	G2	G2	G2	G1	G3	G2	G2
	BA	G2	G2	G2	G2	G2	G2	G1	G3	G2	G2
Amino acid	NJ	G2	G2	G2	G2	G2	G2	G3	G2	G2	G2
	ML	G2	G2	G2	G2	G2	G2	G3	G2	G2	G2
	BA	G2	G2	G2	G2	G2	G2	G3	G3	G2	G3

[1]Non-structural gene, Orf1a+Orf1b
[2]Structural gene, S+E+M+N

than 900 out of 1,000 on bootstrap sampling. Similar results were seen with the BA trees.

As the primary objective of this study was to gain increased insight into the evolutionary relationship between SARS-CoV and other coronavirus groups, we attempted to determine which of the coronavirus groups shared the most recent common ancestor with SARS-CoV in the phylogenetic trees. Although the majority (44 out of 54) of the topologies of the trees pointed to G2-CoV as the nearest neighbor of SARS-CoV (Table 2), some exceptions were noted. When the nucleotide sequences of the E gene were utilized in the construction of the phylogenetic trees, G1-CoV appeared to represent the nearest neighbor of SARS-CoV, by any of the tree-generating techniques. However, phylogenetic trees predicated on amino acid sequences

suggested that avian G3-CoV seemed to be the nearest neighbor to SARS-CoV (Table 2). In the case of the M gene, the situation was even more complex. Whereas G3-CoV appeared to be the nearest neighbor of SARS-CoV on all of the M gene trees predicated on the nucleotide sequences, completely different tree topologies were observed, according to the tree-generating methods utilized, when the amino acid sequences of the M gene were the subject of analysis. The NJ- and ML-based M protein trees exhibited identical topology, in that SARS-CoV appeared to be the nearest neighbor to G2-CoV (Table 2). However, G3-CoV appeared to be the variant most closely related to SARS-CoV on the M protein trees constructed via the BA method.

Therefore, phylogenetic trees constructed based on the nucleotide or amino acid sequences of the Orf1a, Orf1b, S and N genes, as well as the full, NS and ST gene concatenated sequences, all bolster the hypothesis that SARS-CoV and G2-CoV may be monophyletic. However, phylogenetic analyses of the E and M genes generated different outcomes, and we theorized that this might have been attributable to the relatively short length of the E and M gene sequences. Thus, we constructed E+M concatenated sequences, and used these sequences to generate a set of phylogenetic trees (Table 2). Surprisingly, the NJ, ML, and BA trees of the E/M-concatenated nucleotide sequences all identified G2-CoV as the nearest neighbor of SARS-CoV, although either G1-CoV or G3-CoV was implicated as the nearest neighbor of SARS-CoV in the E gene nucleotide trees or M gene nucleotide trees, respectively. In a similar fashion, G2-CoV was implicated as being the most closely related to SARS-CoV in the protein trees of the E/M concatenated sequences, when NJ or ML methods were utilized for analysis (Table 2). The exception to this was the E/M concatenated sequence tree generated by the BA method, in that G3-CoV was determined to be the nearest neighbor to SARS-CoV in the E, M or E/M concatenated sequences.

Similarity analysis of the nucleotide and amino acid sequences

The nucleotide and amino acid sequences of the SARS-CoV genes were compared with those of the other coronavirus groups, and the inter-group (between SARS-CoV and other groups) sequence similarities were determined. Sequence similarities were assessed for each of the genes and concatenated sequences, and expressed as a percentage of similarity (Fig. 2). Non-structural genes were generally determined to exhibit a lesser degree of variation than was seen in conjunction with structural genes. Among the individual genes, Orf1b exhibited the lowest variance, and the E gene evidenced the most profound degree of variance.

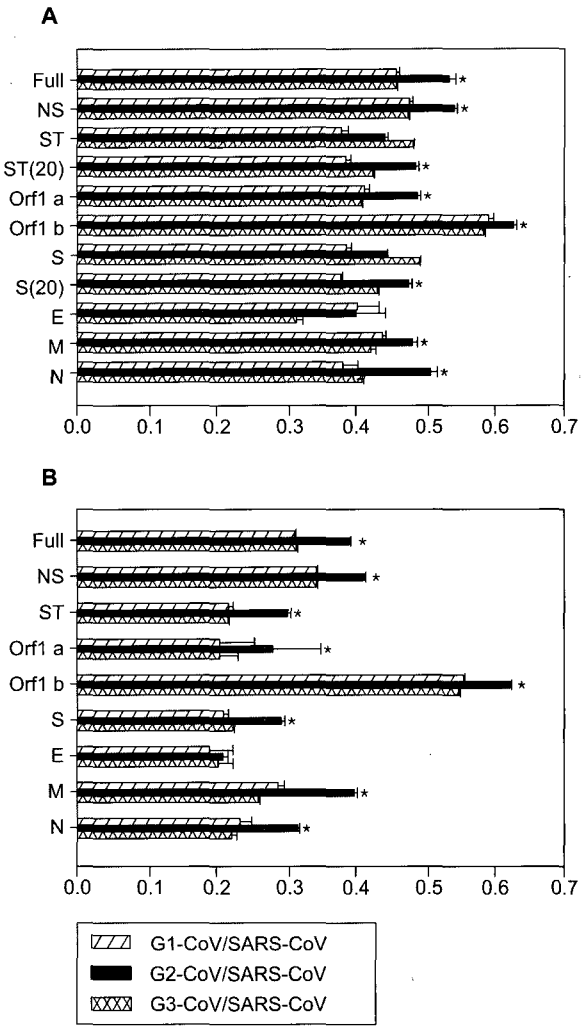

Fig. 2. Sequence similarity between SARS-CoV and other coronavirus groups. Nucleotide and amino acid sequences of coronavirus genes were multiple-aligned using ClustalX. Aligned sequences were compared with SeqAid (ver 0.91) to obtain similarity values among sequences belonging to two different coronavirus groups. Then, mean and standard error of the mean (SEM) values for each pair of comparisons were calculated, and the results were plotted on a bar-graph, using SigmaPlot (ver. 8.0). The asterisks indicate that the similarity between SARS-CoV and G2-CoV is significantly greater than the similarity between SARS-CoV and G1- or G3-CoV, with p values of less than 0.05. A, nucleotide sequence similarity; B, amino acid sequence similarity.

The nucleotide sequence similarity between SARS-CoV and other coronavirus groups suggested that the SARS-CoV sequences were most similar to G2-CoV ($p<0.001$), with the exception of the E and M genes (Fig. 2A). The E gene sequence of SARS-CoV was found to be more similar to G1-CoV or G3-CoV than G2-CoV ($p<0.001$), and the M gene sequence of

SARS-CoV was found to be more similar to G3-CoV than to G1- or G2-CoV ($p<0.001$). When the E+M gene concatenated sequences were taken into consideration, however, SARS-CoV was found to be more similar to G2-CoV than to G1-CoV or G3-CoV ($p<0.001$).

The overall degree of amino acid similarities between SARS-CoV and the other coronavirus groups were less profound than the nucleotide similarities (compare Fig. 2A and Fig. 2B). Nonetheless, with the exception of the E gene, the inter-group amino acid sequence similarities between SARS-CoV and G2-CoV were, again, significantly more similar than those of G1- or G3-CoV ($p<0.001$) (Fig. 2B). In the case of the E gene, none of the coronavirus group members were relatively similar to SARS-CoV ($p>0.05$).

Analysis of the protein secondary structures

The HNN (Hierarchical Neural Network) method, a component of the ExPASy package, allows for the prediction of secondary structures from a given amino acid sequence. This technique predicts the relative percentages of each secondary structure, including alpha helices, extended strands, and random coils, as well as their relative positions within the primary structure. As the latter parameter does not readily lend itself to quantitative analysis, the relative percentages of the secondary structures predicted by HNN for the coronavirus genes were used for purposes of comparison. Data were obtained for the Orf1a, Orf1b, S, E, M, and N genes, and none of the concatenated sequences were included in this study, as the structures of the concatenated proteins were not able to be determined. For easy visual comparison, the average percentage number of each secondary structure of a given SARS-CoV gene was plotted against that of the corresponding gene in the other coronavirus group representatives. As there were 3 secondary structures (alpha-helix, extended strand, and random coil) and 6 genes (Orf1a, Orf1b, S, E, M, and N), each of the scattergrams harbored 18 parameters or points (Fig. 3). Among these 18 parameters, 17 were found to be statistically related upon the comparison of SARS-CoV and G2-CoV, whereas only 7 and 10 were determined to be statistically significantly related upon the comparison of SARS-CoV and G1-CoV or G3-CoV, respectively (data not shown). When the data points were plotted on a scattergram, the strongest correlation was detected between SARS-CoV and G2-CoV, at an r^2 of 0.970, whereas the correlation between SARS-CoV and G3-CoV was the lowest, with an r^2 of 0.839 (Fig. 3). The equation of the regression line for the SARS-CoV and G2-CoV pair was: $Y = 0.99X + 0.42$, which suggests that the secondary

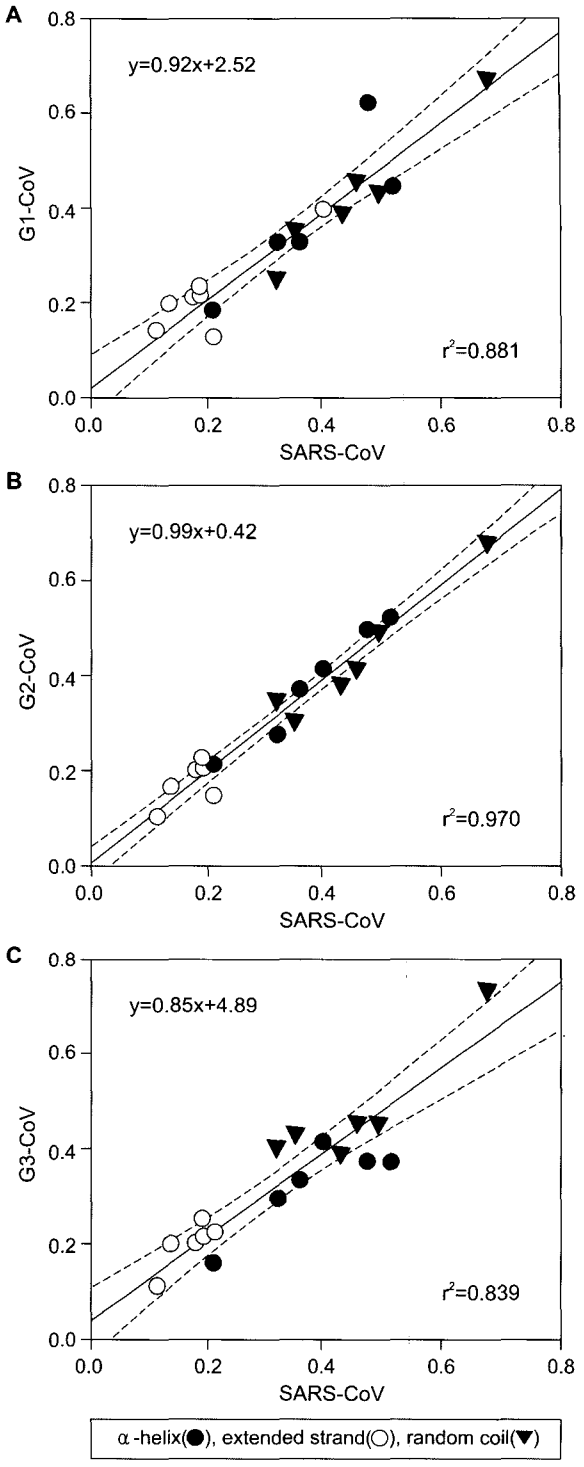

Fig. 3. Relative similarity of the secondary structures of SARS-CoV and other coronavirus groups. Secondary, structures, including alpha helices, extended strands, and random coils of coronavirus proteins, were predicted using the HNN program in the ExPASy package. The relative proportions of the 3 secondary structures of the SARS-CoV proteins were plotted against those of the G1-, G2- or G3-CoV proteins, using SigmaPlot. Regression analyses were conducted for each of the scattergrams, and a regression line with an equation, 95% confidence interval lines, and r^2 (regression coefficient) are shown in the scattergram.

structures of SARS-CoV and G2-CoV are almost identical in composition. Therefore, the predicted secondary structures of the non-structural and structural genes of the coronaviruses appear to implicate G2-CoV as the closest relative of SARS-CoV.

Discussion

The extensive molecular phylogenetic and comparative genomics analyses of the coronavirus genomes described in this study, in aggregate, appear to suggest that SARS-CoV and G2-CoV are monophyletic, and that SARS-CoV is evolutionary related more closely to G2-CoV than to G1- or G3-CoV. Our data are generally consistent with the findings of previous reports (Ksiazeck *et al.*, 2003; Marra *et al.*, 2003; Rota *et al.*, 2003; Zeng *et al.*, 2003; Gibbs *et al.*, 2004; Lio and Goldman, 2004; Zhu and Chen, 2004), but intimations and hypotheses regarding recombination or genomic mosaics (Rest and Mindell, 2003; Stanhope *et al.*, 2004; Stavrinides and Guttman, 2004) were not particularly supported by our data.

Phylogenetic trees constructed on the bases of nucleotide or amino acid sequences allow us to determine how closely or distantly the specific sequence of interest is related to a given known sequence. Thus, in this study we have constructed 60 phylogenetic trees, using the nucleotide or amino acid sequences of 10 different regions from coronavirus genomes, using 3 different techniques. Regardless of the method utilized for tree construction, the SARS-CoVs were found to have clustered into a discrete group, which was readily and obviously separable from the 3 previously characterized coronavirus groups. Thus, coupled with the findings of previous reports (Ksiazeck *et al.*, 2003; Marra *et al.*, 2003), our results indicate that the SARS-CoVs constitute a 4th group of the genus *Coronavirus*. The results of our tree topological analyses revealed that the SARS-CoV and group 2 coronavirus (G2-CoV) were monophyletic in the majority (49 out of 60 trees) of cases. The separation between SARS CoV and G3-CoV was more evident in the trees generated in the present study than in any of the trees generated in previous reports, in which SARS CoV and IBV were either minimally separated by very short branches, or were joined artificially at deep branches (Drosten *et al.*, 2003; Marra *et al.*, 2003; Rota *et al.*, 2003). The exceptions to this phenomenon were the E and M genes, which were found to be smaller than the other coronavirus genes. In the trees based on the E genes, SARS-CoV appeared to be monophyletic with G1-CoV (nucleotide sequence) and G3-CoV (amino acid sequence), regardless of the tree-generating methods employed. In the M gene trees constructed on the basis of nucleotide sequence, SARS-CoV was determined to be monophyletic with G3-CoV. These results are not exclusive to our study, as earlier studies conducted using M gene sequences suggested that SARS-CoV might be monophyletic with G3-CoV (Marra *et al.*, 2003), thus shedding no light on the phylogenetic relationship between SARS-CoV and the other coronavirus groups (Rota *et al.*, 2003; Zeng *et al.*, 2003) whereas phylogenetic analyses with longer sequences, including the Orf1a, Orf1b, S, and N genes, suggested that SARS-CoV and G2-CoV are monophyletic. Our exceptional results with the E and M genes may be attributed to intrinsic genetic properties, or to the lack of sufficient phylogenetic information, due to the smaller size of these genes. The E and M genes of SARS-CoV are 231 and 666 nucleotides in length, respectively, whereas the S and N genes encompass 3,768 and 1,269 nucleotides, respectively. Thus, the E and M genes harbor only approximately 18% and 52% of the information contained in the N gene. In fact, the bootstrap value for the clustering of SARS-CoV and G3-CoV in E protein NJ tree was less than 70%, or 536 out of 1,000. Therefore, the clustering of SARS-CoV and G3-CoV does not appear to be very strong for the E gene. The concatenation of the E and M genes resulted in an increase in the gene size, as well as the amount of available phylogenetic information. The phylogenetic trees constructed on the basis of the E/M gene concatenated sequences implicated SARS-CoV as the nearest neighbor to G2-CoV, thereby suggesting that smaller gene size of the E and M genes may have been the cause of their remarkable tree topologies.

The hypothesis of the mosaic nature of the SARS-CoV genome, as advanced by Stavrinides and Guttman (2004) and Stanhope *et al.* (2004), is predicated on the notion of a previous recombination event between a mammalian coronavirus and an avian coronavirus. This hypothesis is supported, to some degree, by our findings. Our phylogenetic data showed that the nonstructural genes (Orf1a, Orf1b) and S genes clearly share a common root with mammalian G2-CoV, whereas the E and M genes, which are located on the 3' end of the SARS-CoV genome, are more or less similar to avian G3-CoV. However, the N gene, which is located at the very end of the 3' region of SARS-CoV, appeared to be basically similar to mammalian G2-CoV. Thus, at least two recombination events may have occurred, one between the S and E genes and one between the M and N genes, thereby supporting the idea of the mosaic origin of the E and M genes of SARS-CoV.

Analyses designed to evaluate the functional relatedness of SARS-CoV genes and other coronavirus groups also indicated that G2-CoV is the most closely

related of the coronaviruses to SARS-CoV. We also conducted analyses of the secondary structures and codon usage characteristics of the coronavirus genes. The data presented in Fig. 3 shows that the composition of the secondary structures in the SARS-CoV proteins was quite consistent with those of G2-CoV, evidencing an r^2 of 0.97. This result does not immediately mandate the idea that the overall tertiary structure of the SARS-CoV proteins is more similar to that of G2-CoV than G1- or G3-CoV, as we were able only to measure the relative proportion, and not the exact position, of each of the secondary structures. Nevertheless, the secondary structure data obtained for the coronavirus proteins should provide some additional support to the notion that SARS-CoV is the coronoviral variant most closely related to G2-CoV.

In conclusion, the results of the extensive phylogenetic and evolutionary analyses conducted in this study suggested that SARS-CoV and G2-CoV are closely related.

Acknowledgement

This work was supported by a grant from the Biodiscovery Research Fund of KISTEP (M1-0311-02 -0001).

References

Drosten, C., S. Gunther, W. Preiser, S. van der Werf, H.R. Brodt, S. Becker, H. Rabenau, M. Panning, L. Kolesnikova, R.A. Fouchier, A. Berger, A.M. Burguiere, J. Cinatl, M. Eickmann, N. Escriou, K. Grywna, S. Kramme, J.C. Manuguerra, S. Muller, V. Rickerts, M. Sturmer, S. Vieth, H.D. Klenk, A.D. Osterhaus, H. Schmitz, and H.W. Doerr. 2003. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* 348, 1967-1976.

Gibbs, A.J., M.J. Gibbs, and J.S. Armstrong. 2004. The phylogeny of SARS coronavirus. *Arch. Virol.* 149, 621-624.

Guan, Y., B.J. Zheng, Y.Q. He, X.L. Liu, Z.X. Zhuang, C.L. Cheung, S.W. Luo, P.H. Li, L.J. Zhang, Y.J. Guan, K.M. Butt, K.L. Wong, K.W. Chan, W. Lim, K.F. Shortridge, K.Y. Yuen, J.S. Peiris, and L.L. Poon. 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* 302, 276-278.

Ksiazek, T.G., D. Erdman, C.S. Goldsmith, S.R. Zaki, T. Peret, S. Emery, S. Tong, C. Urbani, J.A. Comer, W. Lim, P.E. Rollin, S.F. Dowell, A.E. Ling, C.D. Humphrey, W.J. Shieh, J. Guarner, C.D. Paddock, P. Rota, B. Fields, J. DeRisi, J.Y. Yang, N. Cox, J.M. Hughes, J.W. LeDuc, W.J. Bellini, and L.J. Anderson; SARS Working Group. 2003. A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* 348, 1953-1966.

Lau SK, P.C. Woo, K.S. Li, Y. Huang, H.W. Tsoi, B.H. Wong, S.S. Wong, S.Y. Leung, K.H. Chan, and K.Y. Yuen. 2005. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl. Acad. Sci. USA.* 102, 14040-14045.

Lio, P. and N. Goldman. 2004. Phylogenomics and bioinformatics of SARS-CoV. *Trends Microbiol.* 12, 106-111.

Marra, M.A., S.J. Jones, C.R. Astell, R.A. Holt, A. Brooks-Wilson, Y.S. Butterfield, J. Khattra, J.K. Asano, S.A. Barber, S.Y. Chan, A. Cloutier, S.M. Coughlin, D. Freeman, N. Girn, O.L. Griffith, S.R. Leach, M. Mayo, H. McDonald, S.B. Montgomery, P.K. Pandoh *et al.* 2003. The Genome sequence of the SARS-associated coronavirus. *Science* 300, 1399-1404.

Peiris, J.S., S.T. Lai, L.L. Poon, Y. Guan, L.Y. Yam, W. Lim, J. Nicholls, W.K. Yee, W.W. Yan, M.T. Cheung, V.C. Cheng, K.H. Chan, D.N. Tsang, R.W. Yung, T.K. Ng, and K.Y. Yuen; SARS study group. 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* 361, 1319-1325.

Poon LL, D.K. Chu, K.H. Chan, O.K. Wong, T.M. Ellis, Y.H. Leung, S.K. Lau, P.C. Woo, K.Y. Suen, K.Y. Yuen, Y. Guan, and J.S. Peiris. 2005. Identification of a novel coronavirus in bats. *J Virol.* 79, 2001-2009.

Rest, J.S. and D.P. Mindell. 2003. SARS associated coronavirus has a recombinant polymerase and coronaviruses have a history of host-shifting. *Infect. Genet. Evol.* 3, 219-225.

Rota, P.A., M.S. Oberste, S.S. Monroe, W.A. Nix, R. Campagnoli, J.P. Icenogle, S. Penaranda, B. Bankamp, K. Maher, M.H. Chen, S. Tong, A. Tamin, L. Lowe, M. Frace, J.L. DeRisi, Q. Chen, D. Wang, D.D. Erdman, T.C. Peret, C. Burns, T.G. Ksiazek, P.E. Rollin, A. Sanchez, S. Liffick, B. Holloway, J. Limor, K. McCaustland, M. Olsen-Rasmussen, R. Fouchier, S. Gunther, A.D. Osterhaus, C. Drosten, M.A. Pallansch, L.J. Anderson, and W.J. Bellini. 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300, 1394-1399.

Snijder, E.J., P.J. Bredenbeek, J.C. Dobbe, V. Thiel, J. Ziebuhr, L.L. Poon, Y. Guan, M. Rozanov, W.J. Spaan, and A.E. Gorbalenya. 2003. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J. Mol. Biol.* 331, 991-1004.

Stanhope, M.J., J.R. Brown, and H. Amrine-Madsen. 2004. Evidence from the evolutionary analysis of nucleotide sequences for a recombinant history of SARS-CoV. *Infect. Genet. Evol.* 4, 15-19.

Stavrinides, J. and D.S. Guttman. 2004. Mosaic evolution of the severe acute respiratory syndrome coronavirus. *J. Virol.* 78, 76-82.

Thiel, V., K.A. Ivanov, A. Putics, T. Hertzig, B. Schelle, S. Bayer, B. Weissbrich, E.J. Snijder, H. Rabenau, H.W. Doerr, A.E. Gorbalenya, and J. Ziebuhr. 2003 Mechanisms and enzymes involved in SARS coronavirus genome expression. *J. Gen. Virol.* 84, 2305-2015.

Yap, Y.L., X.W. Zhang, and A. Danchin. 2003. Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. *BMC Bioinformatics* 4, 43.

Zeng, F.Y., C.W. Chan, M.N. han, J.D. Chen, K.Y. Chow, C.C. Hon, K.H. Hui, J. Li, V.Y. Li, C.Y. Wang, P.Y. Wang, Y. Guan, B. Zheng, L.L. Poon, K.H. Chan, K.Y. Yuen, J.S.

Peiris, and F.C. Leung. 2003. The complete genome sequence of severe acute respiratory syndrome coronavirus strain HKU-39849 (HK-39). *Exp. Biol. Med (Maywood)*. 228. 866-873.

Zhu, G. and H.W. Chen. 2004. Monophyletic relationship between severe acute respiratory syndrome coronavirus and group 2 coronaviruses. *J. Infect. Dis*. 189, 1676-1678.