JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 12, Number 8, 2005
© Mary Ann Liebert, Inc.
Pp. 1103–1116

Genomic Classification Using an Information-Based Similarity Index: Application to the SARS Coronavirus

ALBERT C.-C. YANG, ARY L. GOLDBERGER, and C.-K. PENG

ABSTRACT

Measures of genetic distance based on alignment methods are confined to studying sequences that are conserved and identifiable in all organisms under study. A number of alignment-free techniques based on either statistical linguistics or information theory have been developed to overcome the limitations of alignment methods. We present a novel alignment-free approach to measuring the similarity among genetic sequences that incorporates elements from both word rank order-frequency statistics and information theory. We first validate this method on the human influenza A viral genomes as well as on the human mitochondrial DNA database. We then apply the method to study the origin of the SARS coronavirus. We find that the majority of the SARS genome is most closely related to group 1 coronaviruses, with smaller regions of matches to sequences from groups 2 and 3. The information based similarity index provides a new tool to measure the similarity between datasets based on their information content and may have a wide range of applications in the large-scale analysis of genomic databases.

Key words: Shannon entropy, SARS coronavirus.

INTRODUCTION

GENETIC DISTANCE MEASURES ARE INDICATORS of similarity among species or populations and are useful for reconstructing phylogenetic relationships (Graur and Li, 1999). Measures of genetic distance are mainly derived from examining each pair of sequences aligned nucleotide-by-nucleotide and estimating the number of substitutions. Since the mechanism of genome evolution relies not only on point-mutations but recombination or horizontal gene transfer from other species, the heterogeneity of gene segments will substantially degrade the accuracy of optimal sequence alignment methods, which are based on the estimation of nucleotide substitution. Therefore, alignment methods are confined to studying sequences that are conserved and identifiable in all organisms under study (Vinga and Almeida, 2003).

Cardiovascular Division and Margret and H.A. Rey Institute for Nonlinear Dynamics in Medicine, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, Massachusetts 02215.

An alternative approach is to develop alignment-free sequence comparison methods. Current alignment-free sequence comparison methods can be classified into two categories (Vinga and Almeida, 2003): information theory-based (Li *et al.*, 2001) and word statistics-based measures (Campbell *et al.*, 1999; Qi *et al.*, 2004; Chaudhuri and Das, 2002; Hao *et al.*, 2003; Karlin and Burge, 1995; Qi *et al.*, 2004; Stuart *et al.*, 2002). We have developed a new index adapted from linguistic analysis and information theory to measure the similarity between symbolic sequences (Yang *et al.,* 2003a, 2003b). Our approach is based on the concept that the information content in any symbolic sequence is primarily determined by the repetitive usage of its basic elements. The novelty of this information-based similarity index is that it incorporates elements of both information-based and word statistics-based categories since the rank order difference of each n-tuple (word statistics) is weighted by its information content using Shannon entropy (information theory) (Shannon, 1948). Furthermore, the composition of these basic elements captures both global information related to usage of repetitive elements in genetic sequences, as well as local sequence order determined by the n-tuple nucleotides. Hence, our method provides a complementary approach to overcoming limitations of alignment methods and is capable of exploring genetic sequences with heterogeneic origins. The resulting measurement has been validated with respect to generic information-carrying symbolic sequences (Yang *et al.*, 2003a, 2003b). Here we show the specific application of this method to genomic sequences.

METHODS

We have recently developed and validated a generic information-based similarity index to quantify the similarity between symbolic sequences. This method, which has been used for analysis of complex physiologic signals (Yang *et al.*, 2003a) and literary texts (Yang *et al.*, 2003b), can be readily adapted to genetic sequences by examining usages of n-tuple nucleotides (“words”). We first determine the frequencies for each n-tuple by applying a sliding window (moving one nucleotide/step) across the entire genome, and then rank each n-tuple according to its frequency in descending order. To compare the similarity between genetic sequences, we plot the rank number of each n-tuple in the first sequence against that of the second sequence. Figure 1 shows the comparison of 4-tuple nucleotide frequencies between the complete mitochondrial genome of two human lineages and those of the human and gorilla. If two sequences are similar in their rank order of n-tuples, the scattered points will be located near the diagonal line (e.g., human versus human). Therefore, the average deviation of these scattered points away from the diagonal line is a measure of the similarity index between these two sequences (Yang *et al.*, 2003a, 2003b).

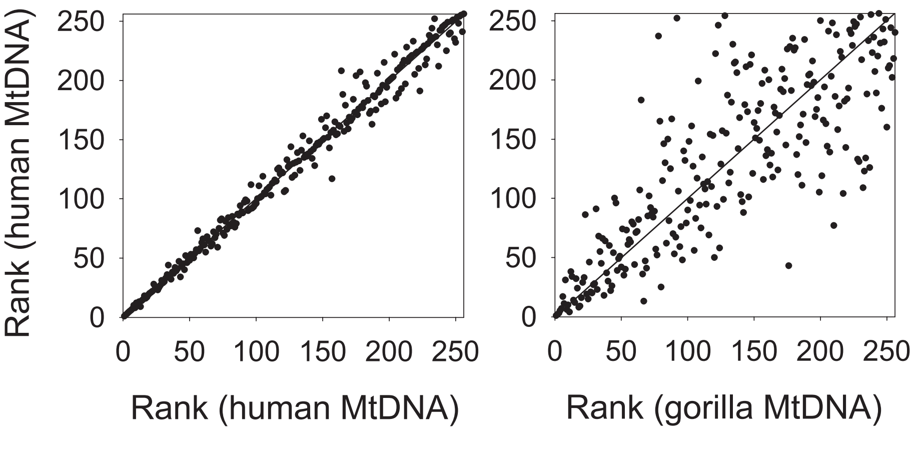


FIG. 1. Rank order comparison of 4-tuple nucleotides (DNA words) of complete mitochondrial DNA (mtDNA) sequences for (**a**) two human lineages and (**b**) human and *Gorilla gorilla*. Words from the two human mtDNA sequences fall close to the diagonal, indicating similar ranking in nucleotide usage. In contrast, the comparison map of human versus gorilla mtDNA yields greater scatter of words around the diagonal. The pairwise distance matrix of virus sequences is then determined (Equation (1)) and used to build a phylogenetic tree using standard distance methods (Saitou and Nei, 1987; Fitch and Margoliash, 1967).

We can define the similarity index (D_n) using n-tuple nucleotides between two sequences, S_1 and S_2, as

$$D_n(S_1, S_2) = \frac{1}{N-1}\sum_{k=1}^{N}|R_1(w_k) - R_2(w_k)|\frac{H_1(w_k) + H_2(w_k)}{\sum_{k=1}^{N}[H_1(w_k) + H_2(w_k)]}. \tag{1}$$

Here $R_1(w_k)$ and $R_2(w_k)$ represent the rank of a specific n-tuple, w_k, in sequences S_1 and S_2, respectively, and $N = 4^n$ is the number of different n-tuple nucleotides. The absolute difference of ranks, $|R_1(w_k) - R_2(w_k)|$, is proportional to the euclidean distance from a given point to the diagonal line. This term is then weighted by the sum of Shannon's entropy H (Shannon, 1948) for w_k in sequences S_1 and S_2. Shannon's entropy measures the information richness of each n-tuple in both sequences. Thus, the more frequently used n-tuples contribute more to measuring similarity among genetic sequences.

We note that this similarity measurement is an empirical index which does not fulfill the criteria of a rigorous distance measure (Yang *et al.*, 2004, 2003b). Therefore, the triangular inequality test is required before generating a phylogenetic tree. When applied to the actual nucleotide sequences here, no violation of the triangular inequality was observed. This similarity metric was then used to determine pairwise distances among genetic sequences and to construct a phylogenetic tree (Felsenstein, 1993; Saitou and Nei, 1987; Fitch and Margoliash, 1967).

To address the statistical reliability of the phylogenetic tree topology, we adapted the methodology of bootstrap analysis (Felsenstein, 1985) and applied it to the information similarity index. Bootstrap analysis is based on the creation of a series of surrogate datasets obtained by resampling the original dataset with replacement. In the case of alignment methods, the surrogate datasets are obtained by resampling aligned columns of nucleotides. To adapt the central concept of bootstrap analysis, we created surrogates by resampling n-tuples from their original distribution in a given sequence. We then calculated the pairwise similarity index between bootstrapped rank-order frequency lists and constructed the phylogenetic tree. The bootstrapped values shown on branches represent the number of successful tests (i.e., those having the same topology as the non-bootstrapped tree) for 1,000 repetitions of the bootstrap procedure.

To further investigate the effects of length of n-tuples on the tree topology, we constructed a mini-database consisting of five known coronaviruses representing three groups. We then estimated the phylogenetic trees based on different lengths of n-tuples ($n = 3$–6). Figure 2a shows the schematic illustration of three established coronavirus groups. Figure 2b–e shows results of neighbor-joining phylogenetic trees based on different lengths of n-tuples ($n = 3$–6). Bootstrapped values (number of successful tests in 1,000 experiments) are shown on the tree branches.

Qualitatively similar results are obtained for n in the range from 3 to 6. The tree with the highest bootstrapped value is obtained by using 4-tuple words. Higher values of n require substantially longer sequences (such that each n–tuple word will be sampled in a statistically meaningful way). As the possible configurations of n-tuples increase exponentially with n, we found that for the coronavirus database (mean sequence length: 29,069 ± 1,569 bp), n values greater than 6 reduce the reliability of the phylogenetic trees.

RESULTS AND DISCUSSION

We first validate the method on the human influenza A viral genomes (Fig. 3) and human mitochondrial DNA database (Fig. 4). The results are comparable to previous reports (Buonagurio *et al.*, 1986; Ingman *et al.*, 2000) addressing the phylogeny of these two databases. We then apply the method to address the origin and classification of the newly identified coronavirus associated with severe acute respiratory syndrome (SARS).

Phylogenetic analysis of the human influenza A virus nonstructural gene database

For initial validation of the information-based similarity index on genetic sequences, we apply this method to the evolution of human influenza A virus from isolates of major outbreaks since 1933 (Buonagurio *et al.*, 1986; Levin *et al.*, 2004). The gene coding for nonstructural (NS) proteins of human influenza A virus

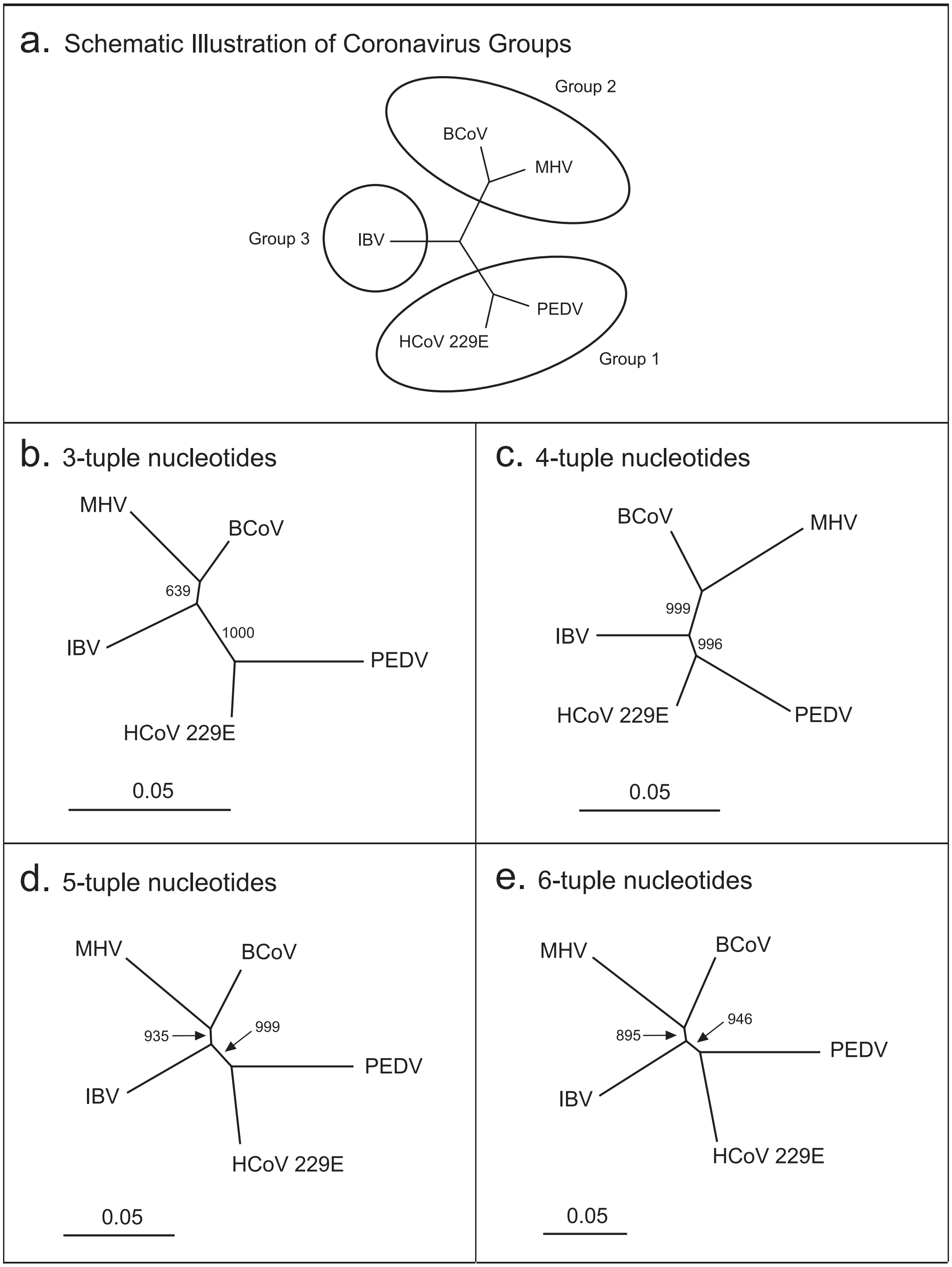


FIG. 2. Comparison of effect of different lengths of *n*-tuples on the topology and statistical significance of phylogenetic trees. Bootstrap values generated by testing 1,000 replicates of the dataset are shown on branches. (**a**) The schematic illustration of three established coronavirus groups. Group 1 (G1): HCoV-229E and PEDV. Group 2 (G2): BCoV and MHV. Group 3 (G3): IBV. (**b**–**e**) Results of neighbor-joining phylogenetic trees based on different lengths of *n*-tuples ($n = 3$–6). Qualitatively similar results are obtained for *n* in the range of 3 to 6. The tree with the highest bootstrapped value is obtained by using 4-tuple words.

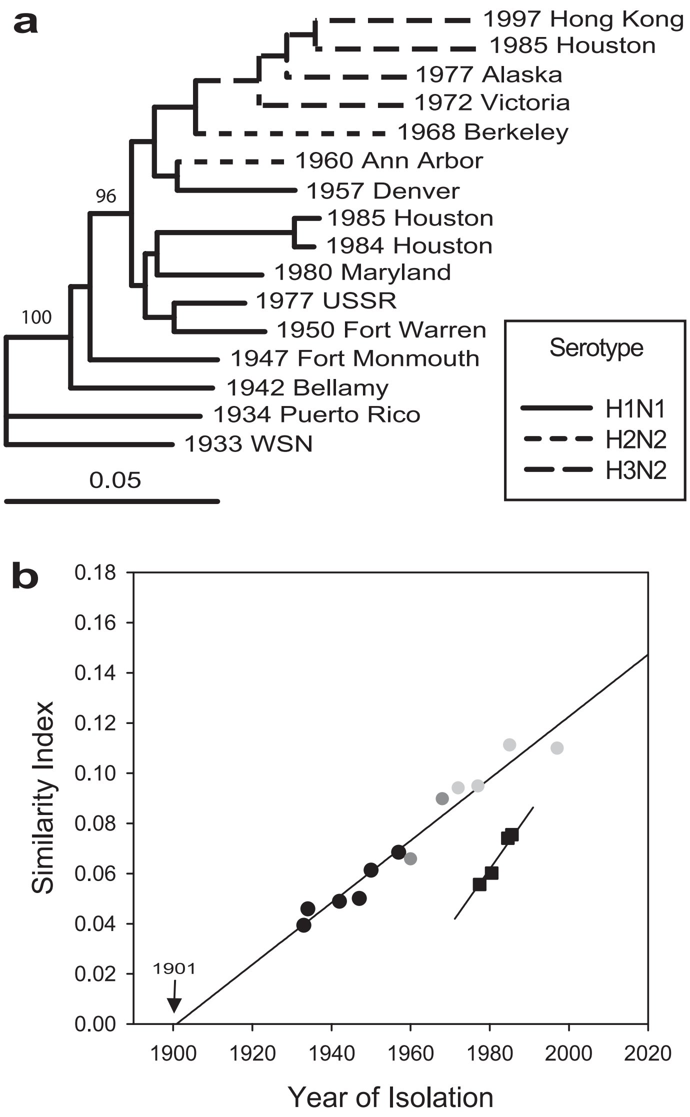


FIG. 3. Evolution of human influenza A virus nonstructural (NS) genes. (**a**) Phylogenetic tree of 16 human influenza A virus NS genes. The number in front of the geographical location indicates the year of the influenza outbreak. Each virus isolate is coded by its hemagglutinin (H) and neuraminidase (N) serotype: H1N1 (solid line), H2N2 (dashed line), and H3N2 (long-dashed line). Each NS gene was analyzed by the method described in the text using 4-tuples as "words." The tree structure is comparable to a prior alignment based study (Buonagurio *et al.*, 1986) showing a rapid and distinct evolutionary path of influenza virus spread during the past 60 years. Bootstrap values generated by testing 100 replicates of the dataset are shown on branches. (**b**) Evolution of the NS gene of 16 human influenza isolates. The graph shows a regression analysis of the year of isolation plotted against the branch length from the common ancestor node at the main trunk of the phylogenetic tree. The figure shows two separated evolutionary pathways (Buonagurio *et al.*, 1986) with a linear correlation between the similarity index with respect to the year of isolation. Furthermore, when analyzing isolates (filled circles) from the major branch of the phylogenetic tree, the year of zero similarity index extrapolated by regression analysis is consistent with the proposed year in which genetic material was introduced from the animal to the human influenza virus (Gammelin *et al.*, 1990). An apparently distinct evolutionary pathway consisting of four H1N1 influenza isolates (77/USSR, 80/Maryland, and 84/85/Houston) is also shown (filled squares).

has demonstrated a rapid and steady mutation rate, making it suitable for studying evolutionary patterns. We collected 16 NS gene sequences which represent isolates from different regions over a span of 60 years (Buonagurio *et al.*, 1986). Each sequence was analyzed by the information-based similarity index using 4-tuples as "words." The neighbor-joining phylogenetic tree is shown in Fig. 3a. By assigning the WSN strain (1933) as the root, the tree shows a progressive evolutionary trend consistent with a previous analysis (Buonagurio *et al.*, 1986). Furthermore, the tree shows distinct evolutionary pathways after the 1947 outbreak. A group of H1N1 subtypes, including 1977 USSR, 1980 Maryland, and 1984/85 Houston isolates, is closely related to the 1950 Fort Warren strain. The others evolved in a separate pathway and had another major genetic shift in 1960 and 1972 which resulted in the H2N2 and H3N2 strains, respectively. These findings are compatible with the consensus view of human influenza A virus evolution (Buonagurio *et al.*, 1986; Levin *et al.*, 2004) and also confirm the unique epidemiology of the H1N1 virus isolated from the USSR in 1977 (Buonagurio *et al.*, 1986).

Since the similarity index proposed here is not based on the assumption that genetic sequences evolve at a constant rate ("molecular clock model") (Graur and Li, 1999), we further investigate the relationship of our similarity measure with respect to the evolutionary time. We calculated the branch length of each isolate from the common ancestor node of the phylogenetic tree shown in Fig. 3a, and plotted against the year of isolation. Figure 3b shows regression analysis of two distinct evolutionary pathways which is consistent with a prior study (Buonagurio *et al.*, 1986). The similarity index of the branch length of each isolate from the common ancestor node indeed linearly correlates with evolutionary time based on the year of isolation. Furthermore, the year of zero similarity index extrapolated by regression analysis is consistent with the proposed year in which genetic material was introduced from the animal to the human influenza virus (Gammelin *et al.*, 1990).

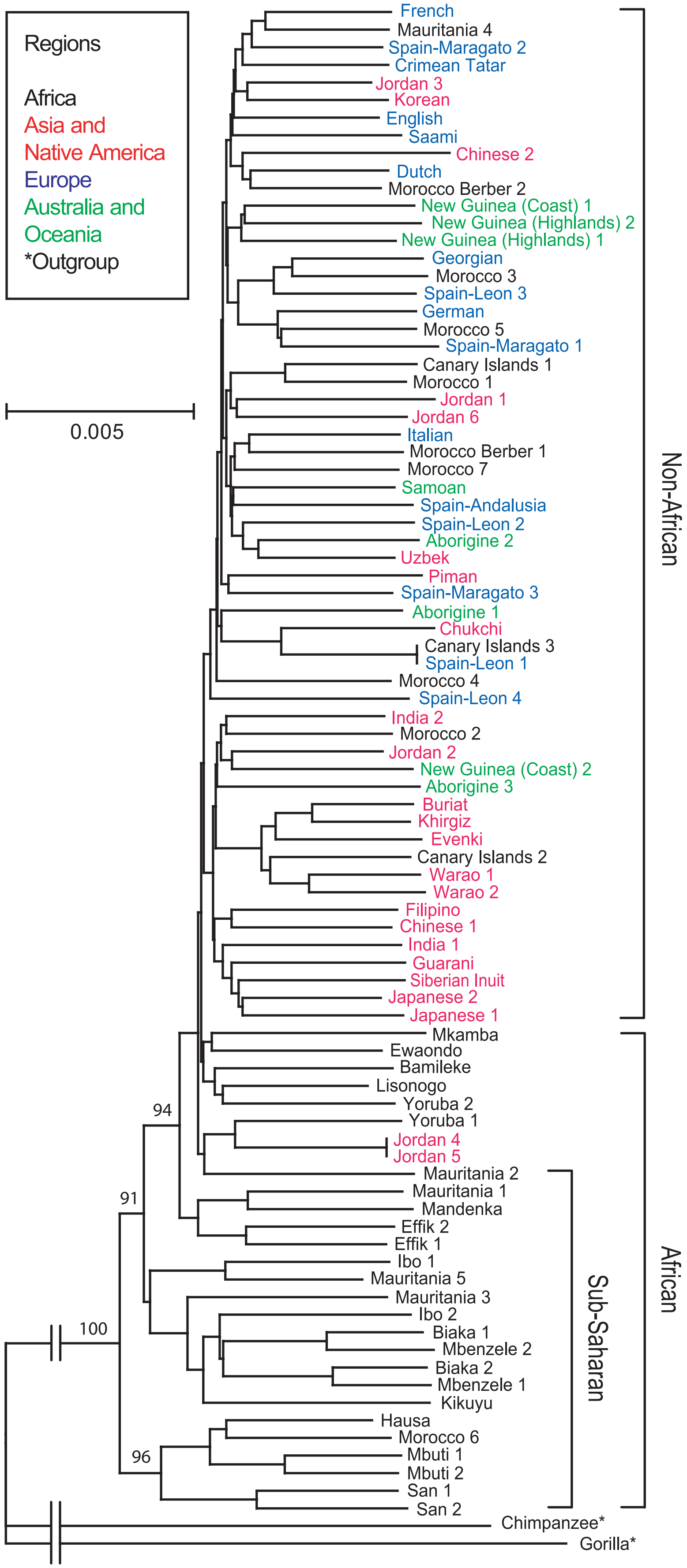
Phylogenetic analysis of the human mitochondrial DNA database

The second part of the validation involves the analysis of complete human mitochondrial DNA (mtDNA) sequences (Cann *et al.*, 1987; Horai *et al.*, 1995; Ruvolo *et al.*, 1993; Vigilant *et al.*, 1991; Ingman *et al.*, 2000; Mishmar *et al.*, 2003). Here we provide an independent analysis without sequence prealignment based on 86 mtDNA sequences (see Appendix for accession numbers) (Ingman *et al.*, 2000; Mishmar *et al.*, 2003). Each sequence was analyzed by the information-based similarity index using n-tuples as "words" ($n = 3$–5). The neighbor-joining phylogenetic tree based on the information-based similarity index is shown in Fig. 4 ($n = 5$). The branching order of each mtDNA sequence is comparable with prior studies based on sequence alignment methods (Horai *et al.*, 1995; Ruvolo *et al.*, 1993; Vigilant *et al.*, 1991; Ingman *et al.*, 2000). All of sub-Saharan African lineages are classified on the bottom of the tree near the root, supporting the African origin of human evolution (Horai *et al.*, 1995; Ruvolo *et al.*, 1993; Vigilant *et al.*, 1991; Ingman *et al.*, 2000). Furthermore, our classification scheme correctly classifies other lineages according to their geographic distribution. For example, Mediterranean people, including Spanish, Italian, and Moroccan lineages, are classified under the same branch and close to the branch of European lineages.

Phylogenetic analysis of the SARS genome

The outbreak of SARS in 2003 has had a tremendous impact on worldwide health care systems (Lee *et al.*, 2003; Poutanen *et al.*, 2003). A central question relevant to the prevention of the recurrence of future SARS outbreak is to determine the virus's origin. Several groups have contributed to identifying and sequencing the complete genome of the newly recognized pathogen, SARS-associated coronavirus (SARS-CoV) (Rota *et al.*, 2003; Marra *et al.*, 2003; Drosten *et al.*, 2003; Peiris *et al.*, 2003). A SARS-like

FIG. 4. The neighbor-joining phylogenetic tree (Saitou and Nei, 1987) based on complete mitochondrial DNA sequences (mtDNA). Eighty-six human mtDNA sequences were available (see Appendix, accession number AF346963-AF347015) (Ingman *et al.*, 2000). We use 5-tuple (or five consecutive nucleotides) as "words." The pairwise distance matrix is calculated using the information-based similarity index. All of the sub-Saharan African lineages are classified at the bottom of tree near the root, which is comparable to prior studies based on sequence alignment algorithms (Horai *et al.*, 1995; Ruvolo *et al.*, 1993; Vigilant *et al.*, 1991) supporting the African origin of human evolution. Bootstrap values generated by testing 100 replicates of the dataset are shown on branches.


Regions
Africa
Asia and
Native America
Europe
Australia and
Oceania
*Outgroup
0.005
French
Mauritania 4
Spain-Maragato 2
Crimean Tatar
Jordan 3
Korean
English
Saami
Chinese 2
Dutch
Morocco Berber 2
New Guinea (Coast) 1
New Guinea (Highlands) 2
New Guinea (Highlands) 1
Georgian
Morocco 3
Spain-Leon 3
German
Morocco 5
Spain-Maragato 1
Canary Islands 1
Morocco 1
Jordan 1
Jordan 6
Italian
Morocco Berber 1
Morocco 7
Samoan
Spain-Andalusia
Spain-Leon 2
Aborigine 2
Uzbek
Piman
Spain-Maragato 3
Aborigine 1
Chukchi
Canary Islands 3
Spain-Leon 1
Morocco 4
Spain-Leon 4
India 2
Morocco 2
Jordan 2
New Guinea (Coast) 2
Aborigine 3
Buriat
Khirgiz
Evenki
Canary Islands 2
Warao 1
Warao 2
Filipino
Chinese 1
India 1
Guarani
Siberian Inuit
Japanese 2
Japanese 1
Mkamba
Ewaondo
Bamileke
Lisonogo
Yoruba 2
Yoruba 1
Jordan 4
Jordan 5
Mauritania 2
Mauritania 1
Mandenka
Effik 2
Effik 1
Ibo 1
Mauritania 5
Mauritania 3
Ibo 2
Biaka 1
Mbenzele 2
Biaka 2
Mbenzele 1
Kikuyu
Hausa
Morocco 6
Mbuti 1
Mbuti 2
San 1
San 2
Chimpanzee*
Gorilla*
94
91
100
96
Non-African
African
Sub-Saharan

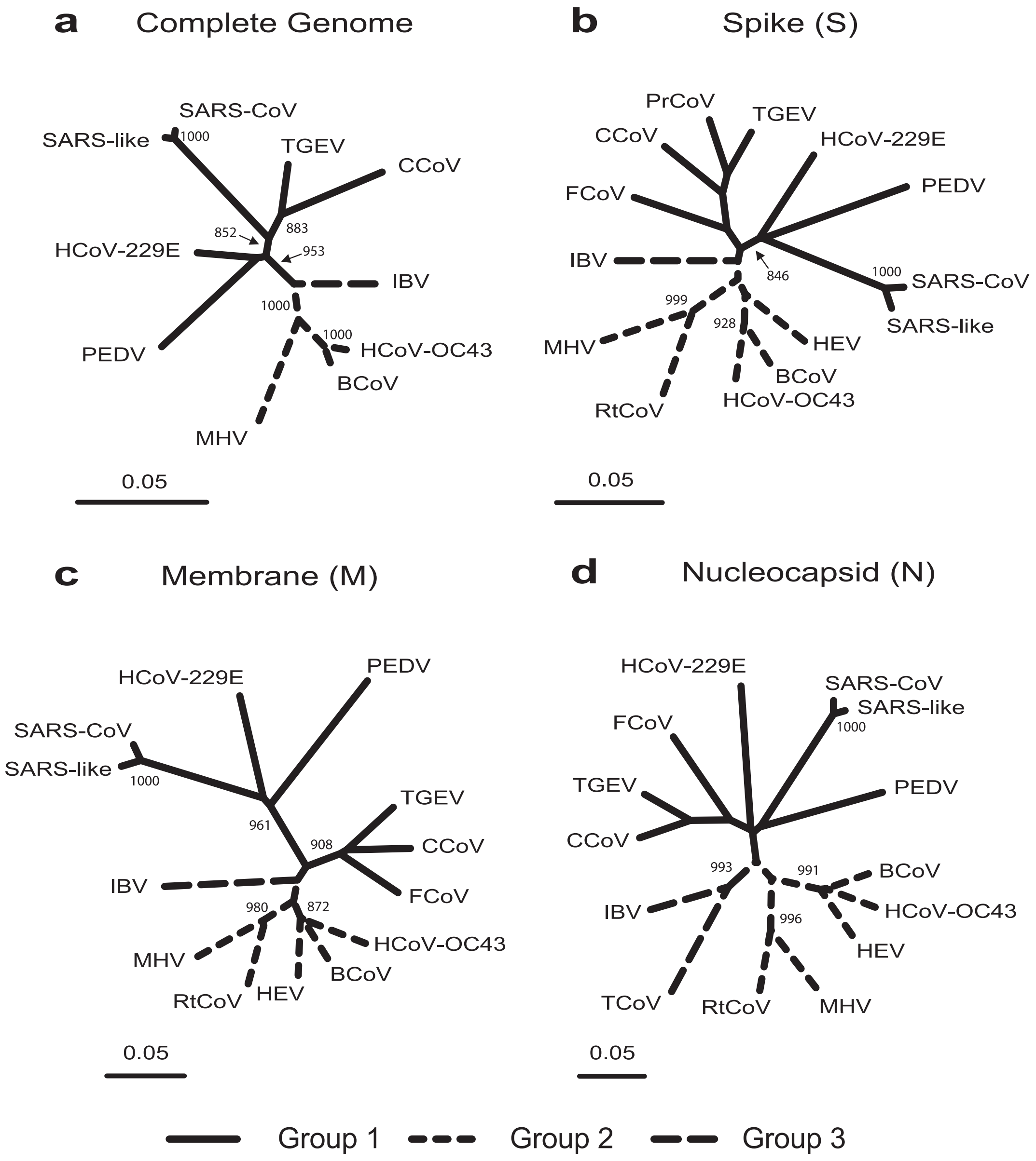


FIG. 5. Neighbor-joining phylogenetic analysis (Saitou and Nei, 1987) of coronaviridae using an information-based similarity index based on 4-tuple nucleotides as "RNA words." (A comparable result is obtained by the Fitch-Margoliash method (Fitch and Margoliash, 1967).) Bootstrap values generated by testing 1,000 replicates of the dataset are shown on branches. (**a**) complete genomes; (**b**) spike glycoprotein; (**c**) membrane protein; and (**d**) nucleocapsid protein. SARS-CoV (NC_00471) and SARS-like virus (AY304486) were compared to the following three major coronavirus groups. Group 1 (G1): human coronavirus 229E (HCoV-229E); transmissible gastroenteritis virus (TGEV); porcine epidemic diarrhea virus (PEDV); canine coronavirus (CCoV); feline infectious peritonitis virus (FCoV); porcine respiratory coronavirus (PrCoV). Group 2 (G2): human coronavirus OC43 (HCoV-OC43); bovine coronavirus (BCoV); porcine hemagglutinating encephalomyelitis virus (HEV); rat sialodacryoadenitis virus (RtCoV); murine hepatitis virus (MHV). Group 3 (G3): avian infectious bronchitis virus (IBV). Comparable results are obtained using different lengths of *n*-tuple words (*n*: 3–5 for the complete genome; *n*: 3–4 for structural protein genes).

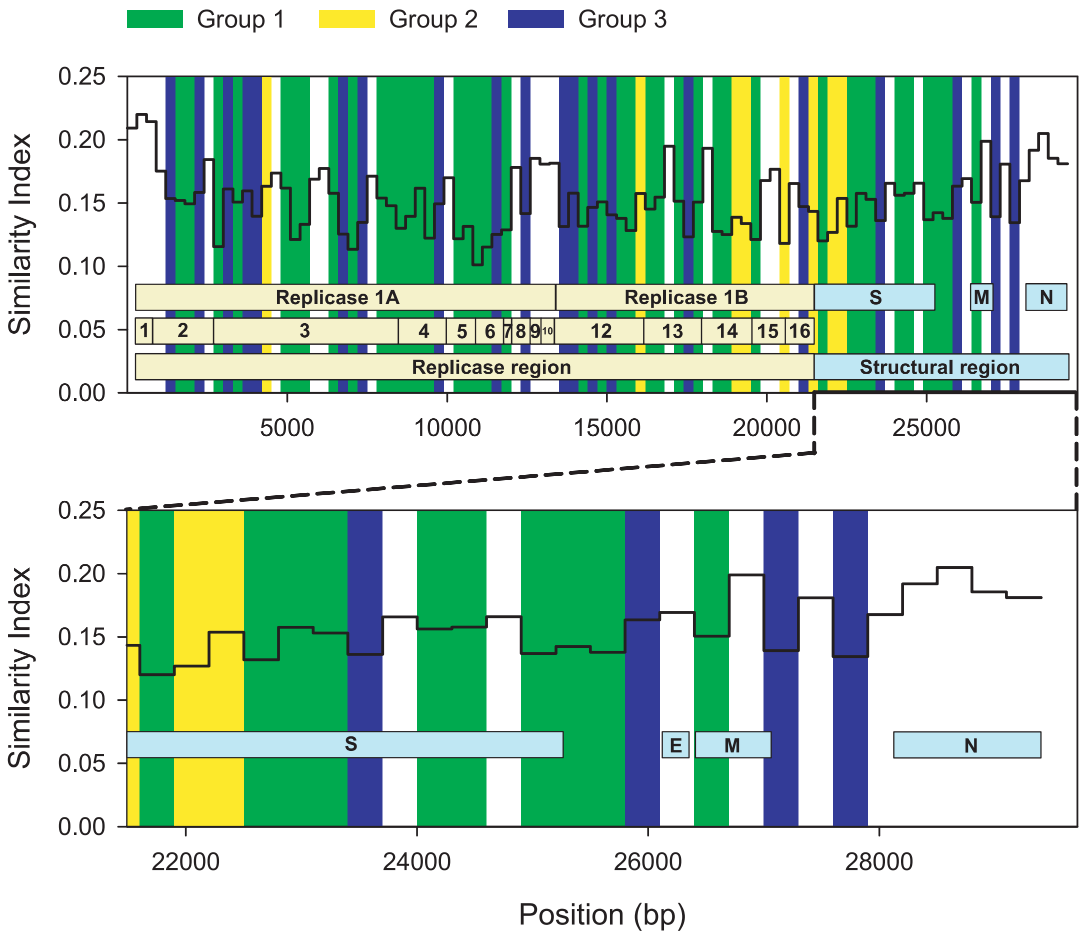


FIG. 6. Comparison of SARS-CoV with the entire genome of other known coronaviruses. The SARS genome is decomposed into nonoverlapping 300 bp segments. Each segment is compared to entire genomes of known coronaviruses to find the best-fit sequence. Each segment is shown by the index of similarity where low values indicate greater similarity to the most closely related coronavirus. To estimate the significance level of the similarity index, we computed the similarity index of 10^5 pairs of randomly selected 300 bp segments from known coronaviruses. We then determined the significance level by computing the 95 percentile rank value of the similarity index (0.165). Only segments within a statistically significant level are color coded to their corresponding group. The white column represents those sequences not significantly similar to any known group. We find that 30% of the entire genome is dis-similar to any known groups. Of note, 41% of the entire SARS-CoV genome is related to group 1 coronaviruses, while only 8% and 21% of the SARS genome are related to group 2 and 3 coronaviruses, respectively. Comparable results are obtained by using a sliding window to decompose the SARS genome.

virus has also been isolated from wild animals such as the palm civet in southern China, indicating that SARS-CoV may have originated from a previously unidentified animal coronavirus (Guan *et al.*, 2003). The relationship of the poorly conserved SARS genome to other coronaviruses, however, is still in question (Vogel, 2003; Enserink and Normile, 2003; Enserink, 2003) since current studies are based on the small portion of aligned sequences (Rota *et al.*, 2003; Eickmann *et al.*, 2003; Marra *et al.*, 2003; Snijder, *et al.*, 2003; Stadler *et al.*, 2003).

Our results based on analysis of available complete coronavirus genomes (Fig. 5a) have some notable distinctions from these previous phylogenetic studies using sequence alignment (Rota *et al.*, 2003; Eickmann *et al.*, 2003; Marra *et al.*, 2003; Stadler *et al.*, 2003). In particular, our method indicates that the SARS-CoV is not classified as a new group but is close to the group 1 coronaviruses. We further analyze the genes coding for structural proteins of SARS-CoV, including the spike glycoprotein (S), the membrane glycoprotein (M), as well as the nucleocapsid protein (N). When examining the phylogeny

Table 1. Homologies between SARS-CoV Genes Other Coronaviruses Using an Information-Based Similarity Index[a]

	Group 1			*Group 2*			*Group 3*
SARS-CoV proteins	*HCoV 229E*	*PEDV*	*TGEV*	*MHV*	*BCoV*	*HCoV OC43*	*IBV*
Nsp1*	0.240	0.232	0.234	0.237	0.267	0.261	**0.209**
Nsp2	0.144	0.139	0.125	0.180	0.167	0.169	**0.115**
Nsp3 (PLpro)	0.094	0.107	**0.073**	0.137	0.129	0.129	0.098
Nsp4	0.113	**0.101**	0.119	0.108	0.116	0.114	0.135
Nsp5 (3CLpro)	**0.093**	0.107	0.102	0.123	0.119	0.121	0.115
Nsp6	0.092	**0.083**	0.089	0.104	0.085	0.083	0.105
Nsp7	0.159	0.168	0.161	0.199	0.179	0.181	**0.155**
Nsp8	0.153	0.156	**0.135**	0.151	0.172	0.163	0.150
Nsp9*	0.198	**0.187**	0.204	0.230	0.215	0.210	0.200
Nsp10*	**0.179**	0.180	0.187	0.210	0.200	0.200	0.188
Nsp12 (RdRp)	**0.081**	0.107	0.088	0.113	0.099	0.095	0.102
Nsp13 (Helicase)	**0.115**	0.117	0.123	0.159	0.150	0.148	0.138
Nsp14	0.097	0.110	**0.094**	0.118	0.113	0.113	0.120
Nsp15	0.128	0.162	0.130	0.154	0.140	0.143	**0.123**
Nsp16	0.114	0.142	**0.107**	0.155	0.146	0.143	0.116
Spike (S)	**0.101**	0.119	0.124	0.128	0.146	0.139	0.143
Envelope (E)*	0.199	0.186	0.184	0.223	0.201	0.182	**0.171**
Membrane (M)	**0.163**	0.196	0.203	0.217	0.219	0.234	0.234
Nucleocapsid (N)*	0.208	**0.165**	0.191	0.209	0.196	0.187	0.204

[a] Numbers indicate the similarity index between individual SARS-CoV genes and other coronaviruses (low value indicates greater similarity). The values of the most closely related coronaviruses are in bold. Asterisks indicate a nonsignificant similarity index (greater than 0.165). Nsp, nonstructral protein. RdRp, RNA-dependent RNA polymerase.

(Figs. 5b, 5c, and 5d) of these three genes, we find that all structural proteins consistently cluster with two viruses: human coronavirus 229E (HCoV-229E) and porcine epidemic diarrhea virus (PEDV). This finding also suggests that SARS-CoV and group 1 coronaviruses are closely related and are likely to share a common ancestor.

Since the SARS genome likely has heterogeneous origins (Rota *et al.*, 2003; Marra *et al.*, 2003), analysis based solely on the whole genome that mixes word distributions from these different sources may not reveal details of segmentary origins. Therefore, we next systemically compare the similarity of individual segments of SARS-CoV to representative coronaviruses from established groups, including HCoV-229E, murine hepatitis virus (MHV), and avian infectious bronchitis virus (IBV). We first decompose the SARS genome into nonoverlapping 300 bp segments. We then compare each segment to the entire genome of other coronaviruses to find the best-fit sequences using the similarity index. Each segment is assigned to the most closely related group (Fig. 6). Only those segments with similarity indexes within a statistically significant range (see Fig. 6 caption) are color coded to the corresponding group. We find that 30% of the entire genome is dissimilar to any known groups. Of note, 41% of the entire SARS-CoV genome is related to group 1 coronaviruses, while only 8% and 21% of the SARS genome are related to group 2 and 3 coronaviruses, respectively. We test the consistency of the results using different segment lengths

Table 2. Top Ranked 4-Tuples of SARS-CoV Genome and Other Coronaviruses[a]

		Group 1			*Group 2*		*Group 3*
Rank	*SARS-CoV*	*HCOV229E*	*PEDV*	*TGEV*	*MHV*	*BCoV*	*IBV*
1	TGCT	**TGTT**	**TGTT**	**TGTT**	**TGTT**	**TGTT**	**TGTT**
2	TGTT	**TTGT**	**TTTT**	**TTGT**	**TTGT**	**TTTT**	**TTGT**
3	ACAA	**TTTG**	**TTGT**	TGGT	**TTTG**	**TTGT**	**TTTT**
4	TTGT	**TTTT**	GTTG	**TTTT**	TTAT	TTTA	**TTTG**
5	AATG	GTTT	TGGT	**TTTG**	TTTA	TTAT	TTAT
6	TACA	GTTG	**TTTG**	**AATG**	**TGTG**	**TTTG**	TTTA
7	TTCT	TGGT	**CTTT**	**ATTG**	**TTTT**	ATTT	GTTT
8	AAAA*	**TGCT**	**TGCT**	TTAT	ATTT	GTTT	TTAA
9	CAAA*	**AATG**	TATG	TTAA	GTTG	TAAT	TGGT
10	TCTT	TTAA	TTAA	TTGA	TATG	TTAA	TAAT
11	ATGT	**TGTG**	TTGG	**AAAA**	**ATGT**	TATT	ATTT
12	AACA	ATTT	GTTT	**ACAA**	GTGT	**TGAT**	**AATT**
13	TGAT	TTAT	**ATGT**	TTTA	TGGT	TATG	TATT
14	TGTG	TTTA	**TGTG**	TGAA	TTAA	TGGT	GTGT
15	TTTT	TTGA	**TGAT**	ATTT	**TGCT**	**ATGT**	GTTA
16	AAAT	**CTTT**	TTGA	**TGCT**	TATT	GTTG	GTTG
17	AATT	**ATGT**	**TCTT**	**TGAT**	GTTT	**TGCT**	**TTCT**
18	ATTG	GTTA	**TTCT**	**AATT**	TAAT	**AATT**	AGTT
19	CTTT	GTGT	GTGT	**ATGT**	**TGAT**	**TGTG**	**TGTG**
20	TTTG	TTGC	TTAT	TATG	GTTA	TAAA	TAAA
Number of 4-tuples shared with SARS-CoV		9	11	12	8	9	7

[a] Twenty top ranked 4-tuples of the entire genome of the SARS-CoV and other coronaviruses. We first determine the frequencies for each 4-tuple by applying a sliding window (moving one nucleotide/step) across the entire genome and then rank each 4-tuple according to its frequency in descending order. Four 4-tuple sequences (TGTT, TTGT, TTTG, and TTTT) shaded gray are common to all viruses in the list. The number of 4-tuple sequences shared with the SARS-CoV (in bold) is higher in group 1 coronaviruses than in other groups. Of interest, in contrast to other known coronaviruses, two 4-tuple sequences (TTTT and TTTG) are less frequent than their complementary sequences (AAAA and CAAA) in the SARS-CoV genome, suggesting an evolutionary pattern that may be related to replication mechanism.

(100, 500, and 1,000 bp). We find that the proportions of attributed origins from the three coronavirus groups are consistent with the results using 300 bp segments.

We also observe that those segments with nonsignificant matches are associated with genes coding for nonstructural proteins (Nsp) 1, 9, and 10, and for structural envelope (E) and N proteins (Table 1). This finding is consistent with prior reports based on sequence alignment analysis (Rota *et al.*, 2003; Marra *et al.*, 2003). However, other sequences including those coding for Nsp 3 and 12 and the S protein show combined origins from other groups. For example, half of S1 domain of the S protein is partly related to the MHV, whereas the remaining sequence is related to HCoV-229E.

Finally, we compare the rank order of specific 4-tuple sequences of the SARS genome with other known coronaviruses (Table 2). Of interest, we find that two sequences—TTTT, and TTTG—are consistently in the top six ranking "words" in non-SARS coronaviruses but in the 15th and 20th rankings in the SARS genome, respectively. Further, in contrast to all other known coronaviruses, the frequencies of these two words in the SARS genome are lower than their complementary sequences (AAAA and CAAA). Whether this n-tuple "word" usage pattern is an incidental finding or is related to the viral evolutionary mechanism remains to be determined.

CONCLUSIONS

In summary, our information-based method provides a complementary approach to study the entire genome as well as individual genes of the SARS coronavirus. The method was originally developed for studying a wide range of symbolic sequences (Yang *et al.*, 2003a, 2003b), and, therefore, assumes minimal knowledge about sequence origin. The method is based on a simple assumption, namely, that genetic sequences from different origins have a preference for certain n-tuples, which they use with higher frequency. Using this method, we have revisited questions concerning the origin of SARS-CoV. Our findings indicate the following key results: 1) the clustering of SARS-CoV, HCoV-229E, and PEDV suggests that SARS-CoV shares a common ancestor with these two viruses which may have exchanged genetic material at an earlier stage; 2) analysis of the entire SARS genome reveals that up to 30% of sequence cannot be reliably attributed to any known coronavirus, in agreement with published studies; and 3) contrary to previous reports (Rota *et al.*, 2003; Marra *et al.*, 2003; Eickmann *et al.*, 2003; Snijder *et al.*, 2003; Stadler *et al.*, 2003), the remainder of the SARS genome is closely related to group 1 coronaviruses, with smaller intermixed segments from groups 2 and 3.

ACKNOWLEDGMENTS

We thank I. Henry, J. Healey, J. Mietus, M. Costa, and S. Izumo for valuable discussions. We gratefully acknowledge support from the NIH/NCRR (P41-RR13622), the NIH/NIA OAIC (P60-AG08812), the James S. McDonnell Foundation, and the G. Harold and Leila Y. Mathers Charitable Foundation.

REFERENCES

Buonagurio, D.A., Nakada, S., Parvin, J.D., Krystal, M., Palese, P., and Fitch, W.M. 1986. Evolution of human influenza A viruses over 50 years: Rapid, uniform rate of change in NS gene. *Science* 232, 980–982.

Campbell, A., Mrazek, J., and Karlin, S. 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondiral DNA. *Proc. Natl. Acad. Sci. USA* 96, 9184–9189.

Cann, R.L., Stoneking, M., and Wilson, A.C. 1987. Mitochondrial DNA and human evolution. *Nature* 325, 31–36.

Chaudhuri, P., and Das, S. 2002. SWORDS: A statistical tool for analysing large DNA sequences. *J. Biosci.* 27, 1–6.

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

Eickmann, M., Becker, S., Klenk, H.D., Doerr, H.W., Stadler, K., Censini, S., Guidotti, S., Masignani, V., Scarselli, M., Mora, M., Donati, C., Han, J.H., Song, H.C., Abrignani, S., Covacci, A., and Rappuoli, R. 2003. Phylogeny of the SARS coronavirus. *Science* 302, 1504–1505.

Enserink, M. 2003. Infectious diseases. Clues to the animal origins of SARS. *Science* 300, 1351.

Enserink, M., and Normile, D. 2003. Infectious diseases. Search for SARS origins stalls. *Science* 302, 766–767.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39, 783–791.

Felsenstein, J. 1993. PHYLIP Phylogeny Inference Package. (3.5c). Department of Genetics, University of Washington, Seattle.

Fitch, W.M., and Margoliash, E. 1967. Construction of phylogenetic trees. *Science* 155, 279–284.

Gammelin, M., Altmuller, A., Reinhardt, U., Mandler, J., Harley, V.R., Hudson, P.J., Fitch, W.M., and Scholtissek, C. 1990. Phylogenetic analysis of nucleoproteins suggests that human influenza A viruses emerged from a 19th-century avian ancestor. *Mol. Biol. Evol.* 7, 194–200.

Graur, D., and Li, W.H. 1999. *Fundamentals of Molecular Evolution*, Sinauer Associates, Boston.

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

Hao, B., Qi, J., and Wang, B. 2003. Prokaryotic phylogeny based on complete genomes without sequence alignment. *Mod. Phys. Lett. B* 17, 91–94.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K., and Takahata, N. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. USA* 92, 532–536.

Ingman, M., Kaessmann, H., Paabo, S., and Gyllensten, U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* 408, 708–713.

Karlin, S., and Burge, C. 1995. Dinucleotide relative abundance extremes: A genomic signature. *Trends Genet.* 11, 283–290.

Lee, N., Hui, D., Wu, A., Chan, P., Cameron, P., Joynt, G.M., Ahuja, A., Yung, M.Y., Leung, C.B., To, K.F., Lui, S.F., Szeto, C.C., Chung, S., and Sung, J.J.Y. 2003. A major outbreak of severe acute respiratory syndrome in Hong Kong. *N. Engl. J. Med.* 348, 1986.

Levin, S.A., Dushoff, J., and Plotkin, J.B. 2004. Evolution and persistence of influenza A and other diseases. *Math. Biosci.* 188, 17–28.

Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., and Zhang, H. 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* 17, 149–154.

Marra, M.A., Jones, S.J.M., Astell, C.R., Holt, R.A., Brooks-Wilson, A., Butterfield, Y.S.N., Khattra, J., Asano, J.K., Barber, S.A., Chan, S.Y., Cloutier, A., Coughlin, S.M., Freeman, D., Girn, N., Griffith, O.L., Leach, S.R., Mayo, M., McDonald, H., Montgomery, S.B., Pandoh, P.K., Petrescu, A.S., Robertson, A.G., Schein, J.E., Siddiqui, A., Smailus, D.E., Stott, J.M., Yang, G.S., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T.F., Bowness, D., Drebot, M., Fernando, L., Flick, R., Garbutt, M., Gray, M., Grolla, A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li, Y., Normand, S., Stroher, U., Tipples, G.A., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R.C., Krajden, M., Petric, M., Skowronski, D.M., Upton, C., and Roper, R.L. 2003. The genome sequence of the SARS-associated coronavirus. *Science* 300, 1399–1404.

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., Sukernik, R.I., Olckers, A., and Wallace, D.C. 2003. Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. USA* 100, 171–176.

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

Poutanen, S.M., Low, D.E., Henry, B., Finkelstein, S., Rose, D., Green, K., Tellier, R., Draker, R., Adachi, D., Ayers, M., Chan, A.K., Skowronski, D.M., Salit, I., Simor, A.E., Slutsky, A.S., Doyle, P.W., Krajden, M., Petric, M., Brunham, R.C., McGeer, A.J., and the National Microbiology Laboratory. 2003. Identification of severe acute respiratory syndrome in Canada. *N. Engl. J. Med.* 348, 1995.

Qi, J., Luo, H., and Hao, B. 2004. CVTree: A phylogenetic tree reconstruction tool based on whole genomes. *Nucl. Acids Res.* 32, W45–W47.

Qi, J., Wang, B., and Hao, B. 2004. Whole genome prokaryote phylogeny without sequence alignment: A K-string composition approach. *J. Mol. Evol* 58, 1–11.

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

Ruvolo, M., Zehr, S., Vondornum, M., Pan, D., Chang, B., and Lin, J. 1993. Mitochondrial COII sequences and modern human origins. *Mol. Biol. Evol.* 10, 1115–1135.

Saitou, N., and Nei, M. 1987. The neighbor-joining method—a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.

Shannon, C.E. 1948. A mathematical theory of communication. *Bell Labs. Tech.* 27, 379–423.

Snijder, E.J., Bredenbeek, P.J., Dobbe, J.C., Thiel, V., Ziebuhr, J., Poon, L.L., Guan, Y., Rozanov, M., Spaan, W.J., and Gorbalenya, A.E. 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.

Stadler, K., Masignani, V., Eickmann, M., Becker, S., Abrignani, S., Klenk, H.D., and Rappuoli, R. 2003. SARS—beginning to understand a new virus. *Nat. Rev. Microbiol.* 1, 209–218.

Stuart, G.W., Moffett, K., and Baker, S. 2002. Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics* 18, 100–108.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., and Wilson, A.C. 1991. African populations and the evolution of human mitochondrial DNA. *Science* 253, 1503–1507.

Vinga, S., and Almeida, J. 2003. Alignment-free sequence comparison—a review. *Bioinformatics* 19, 513–523.

Vogel, G. 2003. SARS outbreak. Flood of sequence data yields clues but few answers. *Science* 300, 1062–1063.

Yang, A.C., Hseu, S.S., Yien, H.W., Goldberger, A.L., and Peng, C.K. 2003a. Linguistic analysis of the human heartbeat using frequency and rank order statistics. *Phys. Rev. Lett.* 90, 108103.

Yang, A.C., Hseu, S.S., Yien, H.W., Goldberger, A.L., and Peng, C.K. 2004. Reply: Comment on linguistic analysis of the human heartbeat using frequency and rank order statistics. *Phys. Rev. Lett.* 92, 109802.

Yang, A.C., Peng, C.-K., Yien, H.-W., and Goldberger, A.L. 2003b. Information categorization approach to literary authorship disputes. *Physica A* 329, 473–483.

Address correspondence to:
C.-K. Peng
Suite KB-28
330 Brookline Ave.
Boston, MA 02215

E-mail: peng@physionet.org

APPENDIX

LIST OF ACCESSION NUMBERS

Coronavirus database

Avian Infectious Bronchitis Virus: NC_001451
Bovine Coronavirus: NC_003045
Canine Coronavirus: AF502583
Feline Infectious Peritonitis Virus: AB086904
Human Coronavirus 229E: NC_002645
Human Coronavirus OC43: M93390
Murine Hepatitis Virus: NC_001846
Porcine Epidemic Diarrhea Virus: NC_003436
Porcine Hemagglutinating Encephalomyelitis Virus: AY078417
Porcine Respiratory Coronavirus: Z24675
Rat Sialodacryoadenitis Virus: AF207551
SARS-CoV: NC_00471
SARS-like virus: AY304486
Transmissible Gastroenteritis Virus: NC_002306

Influenza database

33/WSN (H1N1): M12597
34/Puerto Rico (H1N1): J02150
42/Bellamy (H1N1): M12596
47/Fort Monmouth (H1N1): K00577
50/Fort Warren (H1N1): K0057
57/Denver (H1N1): M12592
60/Ann Arbor (H2N2): M12591
68/Berkeley (H2N2): M12590
72/Victoria (H3N2): AY210316
77/Alaska (H3N2): K01332
77/USSR (H1N1): K00578
80/Maryland (H1N1): M12595
84/Houston (H1N1): M12594
85/Houston (H1N1): M12593
85/Houston (H3N2): M17699
97/Hong Kong (H3N2): AF256183

Mitochondrial DNA database

Gorilla gorilla (Gorilla): X93347
Homo sapiens (Human): AF346963-AF347015
Pan troglodytes (Chimpanzee): X93335