Classifying SARS-CoVs Based on Their Unique Sequences*

Pang-Hao Liu[1] **Sheng-Lung Peng**[1†] **Chung Yi Tang**[2] **Yu-Wei Tsay**[1] **Jason T. L. Wang**[3]

[1] Department of Computer Science and Information Engineering,
National Dong Hwa University, Hualien, Taiwan, R.O.C.
[2] Department of Computer Science,
National Tsing Hua University, Hsinchu, Taiwan, R.O.C
[3] Graduate Program in Computational Biology,
New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA

Abstract

In this paper we propose a method to construct a phylogenetic tree for SARS-CoVs according to their unique sequences. The unique sequences are composed of simple motifs that specify the most representative patterns among species [4]. We propose a new distance measure for building a tree which is independent of any evolutionary model. With the distance measure, the tree is constructed without replying on multiple sequence alignment. Experimental results show that 98% of our clustering groups are similar to the result of ClustalW (Thompson *et al.*, 1994) with support and confidence being higher than 85%. Moreover, we observe that the unique sequences are highly conservative in SARS-CoVs.

Keywords: SARS-CoV, phylogenetic tree, unique sequence, host-shifting.

1. Introduction

In the post-genomics era, a variety of living organisms are sequenced and annotated by laboratories. As a result, a large number of genome databases are developed. The need to efficiently mine the databases to find significant information is urgent and great. Many researches focus on finding repeated sequences and signatures [1–3,5]. In this paper we propose a method for classifying coronavirus by using a particular set of short sequences, namely, unique sequences, in the genomes. These unique sequences are composed of simple motifs that specify the most representative patterns among species.

The coronaviruses of family Coronaviridae are large, enveloped positive stranded RNA viruses that cause respiratory and enteric diseases [9]. The length of their genome sequences is about 30,000 nucleotides. They are the longest RNA viruses found in any of the existing RNA viruses. There are three groups of coronaviruses—groups I and II contain mammalian viruses, whereas group III contains only avian viruses. Within each group, the coronaviruses are classified into distinct species by host ranges. During the recent SARS (severe acute respiratory syndrome) outbreak, there is an evidence showing that the etiologic agent of SARS is a new coronavirus [6], referred as SARS-CoV.

The molecular and biological characteristics of SARS-CoVs are demonstrated in recent studies [4,6–10]. Most of the studies are focused on the classification of SARS-CoVs among Homo sapiens and other animals. There is a considerable evidence showing that the coronaviruses have a history of host-shifting [8]. In the majority of its genomic regions, SARS-CoV is closely related to one of the groups among these regions [4]. In our study, we find that there almost exist six unique sequences in each of the 100 SARS viruses. According to their locations, we propose a method to classify SARS-CoVs and then produce an evolutionary tree for these SARS viruses. Most of the existing tree construction methods use multiple sequence alignment and assume some kind of evolutionary models, such as parsimony or maximal likelihood. By contrast, our method is based on the relative information between the sequence diversities. We suspect that these highly conserved unique sequences among the SARS genomes preserve some major functions.

2. Preliminaries

A subsequence s is *unique* in a sequence S if s appears only once in S and no substring of s appears only once in S. Table 1 gives examples of unique sequences in three species of the bacteria Pseudomonas. In the table, the attribute 'Acc Number' represents the accession number of a species in GenBank; 'Species' denotes the scientific name of a species; 'Length (bps)' denotes the length of the complete sequence of a species; 'Uni-Length' is the minimal length of the unique sequences in the species; 'Num-Seq' denotes the total number of unique sequences with the minimal length in a species; 'Sequence' shows the first unique sequence in the species; 'Position' denotes the starting position

*Supported in part by the National Science Council under grant NSC 90-2213-E-259-014.
†Author to whom correspondence should be addressed.
E-mail: lung@csie.ndhu.edu.tw

of the first unique sequence in the species.

For example, the accession number of Pseudomonas aeruginosa is NC_002516. The total length of this species is 6,264,403 basepairs. The minimal length of the unique sequences in the species is 8. The total number of the unique sequences with length 8 is 92. The first unique sequence with length 8 starts at the position 4,180,365 of the species.

3. Main results

We explore the 100 SARS genomes to find unique sequences of each SARS-CoV. We obtain the following information. All SARS viruses almost have six common unique sequences. They are 'CGGGC,' 'TCCCG,' 'CCCGG,' 'CCCCT,' 'CGCCT,' and 'CCGGA.' According to their positions in sequences, we can divide 100 SARS viruses into 19 groups. Figure 1 shows the classification rules. Note that Figure 1 only describes the relationship among the first 17 groups. Two other two groups, i.e., 18 and 19, are not involved because their characteristics are far away from others.

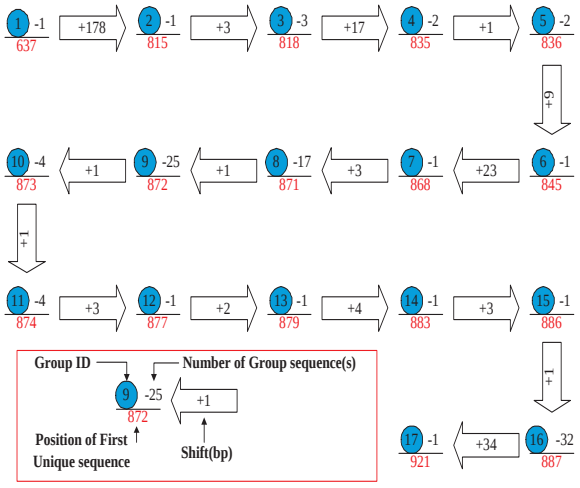

Figure 1. Our classification rules.

Table 2 shows the 19 clusters and the positions of the six unique sequences in their complete sequences.

According to our classification, we discover that it tallied with phylogenetic relationship compared with the method of multiple sequence alignment. Two examples shown in Figures 2 and 3 give an evidence that our classification is similar to other research. Figure 2 comes from *National Center for High-performance Computing* and Figure 3 is depicted by Qin *et al.* [7]. In these figures, we know that the three SARS viruses classified in Group 16 are much closed to each other. These five SARS viruses can be separated into two parts and three branches according to our rules.

As shown in Figure 4 [10], the phylogenetic tree is obtained by applying PAUP* (maximum-likelihood methods using the phylogeny distance model). Most of the clusters are well located.

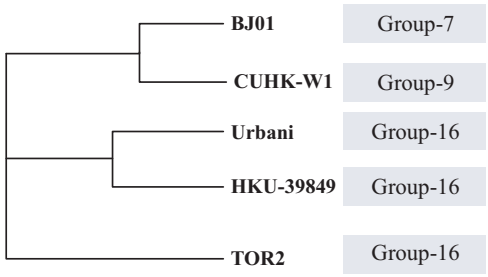

Figure 2. A phylogenetic tree proposed by National Center for High-performance Computing.

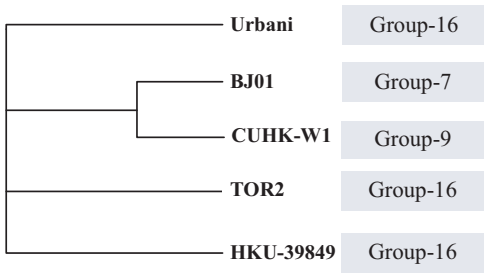

Figure 3. Another phylogenetic tree proposed in [7].

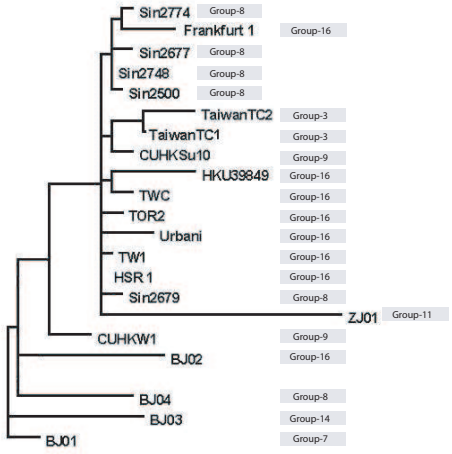

Figure 4. A phylogenetic tree proposed in [10].

We propose a phylogenetic tree for the 100 SARS-CoVs with a 2-layer architecture. The first layer indicates the major classification of all the 100 SARS-CoVs. The second layer indicates the detailed categories within a group. Since the positions of the six unique sequences are the same in a group, we use multiple sequence alignment algorithm to distinguish them. A complete discussion about the tree topology will not be presented in this paper because of the limitation of pages. Figure 5 shows our results. For example, Group

Acc.Number	Species	Length(bps)	Uni-Length	Num-Seq	Sequence	Position
NC_002516	Pseudomonas aeruginosa	6264403	8.	92	AATTAGAT	4180365
NC_002947	Pseudomonas putida	6181863	8.	90	AACTAGTA	1208751
NC_004578	Pseudomonas syringae	6397126	8.	39	AATCGAAC	1005002

Table 1. Unique sequences in three bacteria Pseudomonas.

Source	strain1	Pos	strain2	Pos	strain3	Pos	strain4	Pos	strain5	Pos	strain6	Pos
Group 1	CGGGC	637	TCCCG	9757	CCCGC	14353	CGCCT	22623	CCGGA	24740	CCGGA	26599
Group 2	CGGGC	815	TCCCG	9935	CCCGC	14532	CGCCT	22801	CCGGA	24918	CCGGA	26777
Group 3	CGGGC	818	TCCCG	9938	CCCGC	14535	CGCCT	22804	CCGGA	24921	CCGGA	26780
Group 4	CGGGC	835	TCCCG	9955	CCCGC	14552	CGCCT	22821	CCGGA	24938	CCGGA	24797
Group 5	CGGGC	836	TCCCG	9956	CCCGC	14553	CGCCT	22822	CCGGA	24939	CCGGA	26798
Group 6	CGGGC	845	TCCCG	9965	CCCGC	14562	CGCCT	22831	CCGGA	24948	CCGGA	24807
Group 7	CGGGC	868	TCCCG	9988	CCCGC	14585	CGCCT	22854	CCGGA	24971	CCGGA	26830
Group 8	CGGGC	871	TCCCG	9991	CCCGC	14588	CGCCT	22875	CCGGA	24974	CCGGA	26833
Group 9	CGGGC	872	TCCCG	9956	CCCGC	14553	CGCCT	22822	CCGGA	24939	CCGGA	26798
Group 10	CGGGC	873	TCCCG	9993	CCCGC	14590	CGCCT	22859	CCGGA	24976	CCGGA	24835
Group 11	CGGGC	874	TCCCG	9994	CCCGC	14591	CGCCT	22860	CCGGA	24977	CCGGA	26836
Group 12	CGGGC	877	TCCCG	9997	CCCGC	14594	CGCCT	22863	CCGGA	24980	CCGGA	26839
Group 13	CGGGC	879	TCCCG	9999	CCCGC	14596	CGCCT	22865	CCGGA	24982	CCGGA	26841
Group 14	CGGGC	883	TCCCG	10003	CCCGC	14600	X	X	CGCCT	24985	CCGGA	26845
Group 15	CGGGC	886	TCCCG	10006	CCCGC	14603	CGCCT	22872	CCGGA	24989	CCGGA	26848
Group 16	CGGGC	887	TCCCG	10007	CCCGC	14604	CGCCT	22873	CCGGA	24990	CCGGA	26849
Group 17	CGGGC	921	TCCCG	10041	CCCGC	14638	CGCCT	22907	CCGGA	25024	CCGGA	26883
Group 18	CGGGC	808	TCCCG	9928	X	X	X	X	CCGGA	24911	X	X
Group 19	CGGGC	887	TCCCG	10019	X	X	CCCCT	22894	X	X	CCGAG	24418

Table 2. The 19 clusters of 100 SARS-CoVs.

3 contains Taiwan TC-1, Taiwan TC-2, and Taiwan TC-3. It is worth noting that 98% of the classifications are consistent with PHYLIP and ClustalW except two viruses.

To show these unique sequences are significant, we examine other coronaviruses which are not SARS-CoVs. As shown in Table 3, most of the unique sequences appearing in SARS-CoVs do not appear in the other coronaviruses. The viruses shown in Table 3 are Murine and Bovine viruses. Neither of them contains the six unique sequences simultaneously. Similarly, Human and Avian coronaviruses also have no such subsequences. It shows that the six unique sequences are highly conservative in SARS-CoVs. As a by-product, we believe that the six unique sequences can be used to determine whether a coronavirus is a SARS-CoV or not.

4. Conclusion

In this paper, we construct a two-layered phylogenetic tree for 100 SARS-CoVs based on the positions of six unique sequences. Unlike many existing phylogeny construction methods, the proposed method does not fully require multiple sequence alignment. Furthermore our distance measure does not use any evolutionary model. We believe that the construction of phylogenetic trees by using unique sequences provides an alternative to estimating distances among particular species, especially viruses with high conservative motif.

The method we proposed can be extended in several ways. A weighted scheme for patterns may be added. A procedure for inspecting other viruses that share the same pattern will be performed. Finding suitable motifs and comparing the unique sequences among related species using different distance measures will be investigated. Finding unique sequences also provides a starting point for drug design, primer design, microarray design, and immunity related problems.

5. References

[1] Y.C. Chang and C.H. Chang, Common Repeat Sequences in Bacterial Genomes, *Journal of Medical and Biological Engineering* **23**(2), 65–72, 2003

[2] A. Fadiel, S. Lithwick, and G. Ganji, Remarkable sequence signatures in archaeal genomes, *Archaea*, **1**, 185–190, 2003.

[3] S. Kurtz and C. Schleiermacher, REPuter: fast computation of maximal repeats in complete genomes, *Bioinformatics* **15**(5), 426–427, 1999.

[4] G. Magiorkinis, E. Magiorkinis, D. Paraskevis, *et al.*, Phylogenetic analysis of the full-length SARS-CoV sequences: evidence for phylogenetic discordance in three genomic regions., *J Med Virol.*, **74**(3), 369–372, 2004.

[5] T.B. Nesterova, S.Y. Slobodyanyuk, E.A. Elisaphenko, *et al.*, Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence, *Genome Res.*, **11**(5), 833–

Other CoV	TCCCG	CCCCT	CCCGG	CCGGA	CGCCT	CGGGC
Murine hepatitis virus strain 2	X	X	X	X	X	X
Murine hepatitis virus strain Penn	X	X	X	X	X	X
Murine hepatitis virus strain ML-10	X	X	X	X	X	X
Bovine coronavirus strain Mebus	X	X	X	X	X	X
Bovine coronavirus strain Quebec	X	X	X	X	X	X
Bovine coronavirus isolate BCoV-LUN	X	X	X	X	X	X
Human coronavirus OC43	X	X	X	O	X	X
Human coronavirus OC43 serotype OC43-Paris	X	X	X	O	X	X
Human coronavirus OC43 strain ATCC VR-759	X	X	X	O	X	X
Porcine epidemic diarrhea virus	X	X	O	O	X	O
Transmissible gastroenteritis virus	X	O	O	X	X	O
Human coronavirus 229E	X	X	X	O	X	O
Avian infectious bronchitis virus isolate BJ	X	X	X	X	X	O
Avian infectious bronchitis virus strain Cal99	X	X	X	X	X	O
Avian infectious bronchitis partridge	X	X	O	X	X	X
Avian infectious bronchitis virus isolate Peafowl	X	X	X	X	X	X

Table 3. A comparison with other coronaviruses using the six unique sequences.

849, 2001.

[6] L.L. Poon, Y. Guan, J.M. Nicholls, *et al.*, The aetiology, origins, and diagnosis of severe acute respiratory syndrome., *Lancet Infect Dis.*, **4**(11), 663–671, 2004.

[7] Qin E'de, *et al.*, A complete sequence and comparative analysis of a SARS- associated virus, *Chinese Science Bulletin*, **48**(10), 941–948, 2003.

[8] J.S. Rest and D.P. Mindell, SARS associated coronavirus has a recombinant polymerase and coronaviruses have a history of host-shifting, *Infect Genet Evolution*, **3**(3), 219–25, 2003.

[9] P.A. Rota and M.S. Oberste, Characterization of a novel coronavirus associated with severe acute respiratory syndrome, *Science* **300**, 1394–1399, 2003.

[10] Vicenzi,E. Canducci,F. Pinna,D. Mancini, *et al.*, Coronaviridae and SARS-associated Coronavirus Strain HSR1, *Emerging Infect. Dis.*, **10**, 413–418, 2004.

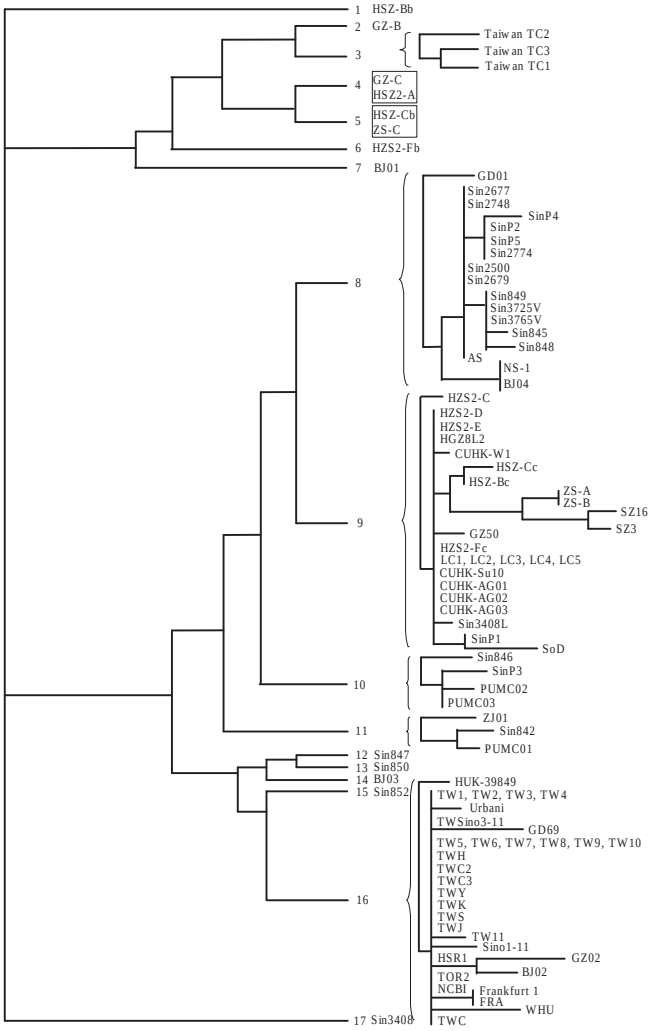

Figure 5. Our phylogenetic tree for SARS-CoVs.