바이오인포매틱스 기법을 활용한 SARS 코로나바이러스의 유전정보 연구

A Study on the Genomic Patterns of SARS coronavirus using Bioinformtaics Techniques

안인성, 정병진*, 손현석*
한국과학기술정보연구원 슈퍼컴퓨팅센터,
서울대학교 보건대학원*

Ahn Insung, Jeong Byeong-Jin*, Son Hyeon S.*
Supercomputing Center, KISTI, Graduate School of Public Health, Seoul National Univ.*

요약

중증급성호흡기증후군(SARS, Severe Acute Respiratory Syndrome)은 전 세계적으로 알려진 바가 없었던 신종 급성 전염성 질환으로써, 2003년 아시아로부터 북미와 유럽지역까지 빠른 속도로 전파되어 나간 이후로부터 많은 과학자들의 연구의 대상이 되어오고 있다. 계통발생학적인 관점에서 SARS 바이러스는 *Coronavirus* 속에 속하는 것으로 알려져 있으나, 전체적인 유전정보 면에서는 다른 코로나바이러스들에 비하여 진화상으로 보존된 부분들이 현저하게 적은 경향을 나타낸다. 자연계에서의 SARS 코로나바이러스(SARS-CoV)의 숙주생물종에 대해서는 아직까지도 명확히 알려지지 않고 있다. 본 연구에서는 SARS-CoV의 유전서열들을 대상으로 다중서열정렬법, 계통발생학적 분석기법 및 다변량 통계분석법 등과 같은 바이오인포매틱스 분석기법들을 활용하여 이 바이러스의 유전정보 패턴을 분석하였다. Relative synonymous codon usage (RSCU)값을 포함하는 여러 유전정보 파라미터들은 *Coronavirus*와 *Lentivirus* 속과 *Orthomyxoviridae* 과로부터 수집된 총 30,305개의 암호화 서열들로부터 계산이 되었으며 이 모든 계산은 KISTI 슈퍼컴퓨팅센터의 SMP 클러스터 상에서 수행되었다. 분석 결과, SARS-CoV는 feline 코로나바이러스와 매우 유사한 RSCU 패턴을 나타내었는데, 이것은 기존에 보고되었던 혈청학적인 연구결과와 일치하는 결과였다. 또한 SARS-CoV와 *human immunodeficiency virus* 및 *influenza A virus*는 공통적으로 각각이 속한 속이나 과 내에서 상대적으로 낮은 RSCU bias를 나타내어서 이와 같은 현상이 이들 바이러스들이 종 간 장벽을 뛰어넘어 전파되는 과정에 영향을 미쳤을 가능성을 시사하였다. 결론적으로 이와 같은 바이오인포매틱스 분석기법들을 활용한 대용량의 유전정보 분석은 유전체 역학 연구에 효과적으로 사용될 수 있을 것으로 기대된다.

Abstract

Since newly emerged disease, the Severe Acute Respiratory Syndrome (SARS), spread from Asia to North America and Europe rapidly in 2003, many researchers have tried to determine where the virus came from. In the phylogenetic point of view, SARS virus has been known to be one of the genus *Coronavirus*, but, the overall conservation of SARS virus sequence was not highly similar to that of known coronaviruses. The natural reservoirs of SARS-CoV are not clearly determined, yet. In the present study, the genomic sequences of SARS-CoV were analyzed by bioinformatics techniques such as multiple sequence alignment and phylogenetic analysis methods as well multivariate statistical analysis. All the calculating processes, including calculations of the relative synonymous codon usage (RSCU) and other genomic parameters using 30,305 coding sequences from the two genera, *Coronavirus*, and *Lentivirus*, and one family, *Orthomyxoviridae*, were performed on SMP cluster in KISTI, Supercomputing Center. As a result, SARS_CoV showed very similar RSCU patterns with feline coronavirus on the both axes of the correspondence analysis, and this result showed more agreeable results with serological results for SARS_CoV than that of phylogenetic result itself. In addition, SARS_CoV, *human immunodeficiency virus*, and *influenza A virus* commonly showed the very low RSCU differences among each synonymous codon group, and this low RSCU bias might provide some advantages for them to be transmitted from other species into human beings more successfully. Large-scale genomic analysis using bioinformatics techniques may be useful in genetic epidemiology field effectively.

I. Introduction

Since newly emerged disease, the Severe Acute Respiratory Syndrome (SARS), spread from Asia to North America and Europe rapidly, many researchers have tried to determine where the virus came from [1]. In 2003, 'the Center for Disease Control and prevention

(CDC)' in Guangdong province analyzed the IgG antibodies of SARS virus from 508 animal traders whose blood was sampled during the outbreak. They resulted that those who handled primarily civets were most likely to have antibodies, followed by traders who dealt in wild boars and muntjac deers [2, 3]. Moreover, Martina *et al.* [4] reported that not only civets, but also other animals might be the reservoirs for SARS virus. They suggested that ferrets and domestic cats were susceptible to SARS infection, and that the virus was efficiently transmitted to animals living with them like market civets.

SARS virus has been known to be one of the genus *Coronavirus*, because it was identified to contain some conserved motifs that were found in other coronaviruses. But, the overall conservation of SARS virus sequence was not highly similar to that of known coronaviruses [5, 6]. Even the two human coronaviruses such as *human coronavirus 229E* and *OC43* which have been known to cause the common cold [7] didn't show any close relationship with *SARS coronavirus* (SARS-CoV). Moreover, SARS-CoV showed no close correlations with other coronavirus species, too [3, 5, 6, 8]. Because of these differences between the serological and phylogenetic analysis, the natural reservoirs of SARS-CoV are not clearly determined, yet.

In this study, we tried to find out some clues, which might determine the emerging characteristics as well as the origins among pandemic RNA viruses including *human immunodeficiency virus* (HIV), *influenza A virus* (Inf_A), and SARS_CoV using genomic patterns. HIV-1 and HIV-2 also thought to be transmitted to the human population through multiple zoonotic infections from non-human primates [9, 10].

Ⅱ. Materials & Methods

1. Genomic Sequences

In order to create the phylogenetic trees for the three major genes including the replicase polyprotein 1ab, spike glycoprotein, and nucleocapsid protein genes, and the whole genome sequences of the genera *Coronavirus*, had been used in this analysis. In the analysis of the synonymous codon usage patterns, 30,305 coding sequences (CDSs) from 38 species of the two family, *Coronaviridae* (1,485 CDSs), and *Orthomyxoviridae* (13,496 CDSs) were used. We also used 13 species of *Lentivirus* genus (15,324 CDSs) which includes HIV-1 and HIV-2 to find out any clue of the genomic patterns among zoonotic-suspected viruses. Using Java codes, all the complete CDSs were extracted from GenBank flat file, and relative synonymous codon usage (RSCU) values of each genus were calculated for the statistical analysis on SMP Cluster system in Korea Institute of Science and Technology Information. Partial CDSs which were not started with the starting codon (AUG), complementary CDSs, and CDSs which had more than one ambiguous sequences were excluded in this study.

2. Phylogenetic Analysis

Multiple sequence alignments were conducted by the ClustalW 1.83 program [11]. Unrooted phylogenetic trees for each gene and whole genome sequence were generated by PAUP*4.0 program [12] using Neighbor-Joining (NJ) method with 1,000 times bootstrapping process. Each gene and its GenBank accession number of SARS-CoVs which was used for phylogenetic analysis is shown in Table 1.

3. Relative Synonymous Codon Usage (RSCU)

RSCU values of each codon were usually used to prevent the amino acid composition influencing the codon usage values of each gene in correspondence analysis (CA) [13]. RSCU values are the number of times that a particular codon is observed, relative to the number of times that the codon would be observed in the absence of any codon usage bias. If there were no codon usage biases, the RSCU value would be 1.00. The RSCU was calculated as

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}} \qquad (1)$$

where X_{ij} was the frequency of occurrence of the jth codon for the ith amino acid, and n_i was the number of codons for the ith amino acid (1). In this study, we used CA method to compare the differences of RSCU values among the species within the two families, *Coronaviridae* and *Orthomyxoviridae*.

4. Statistical Analysis

Correspondence analysis (CA) is a graphical procedure for representing associations in a table of frequencies or counts. If the contingency table has *I* rows and *J* columns, the plot produced by CA contains two sets of points: A set of *I* points corresponding to the rows and a set of *J* points corresponding to the columns. The positions of the points reflect associations. RSCU values on 59 codons, which were described above, were used, and all the CA process was performed by the statistical program, SAS program 9.1 release [14].

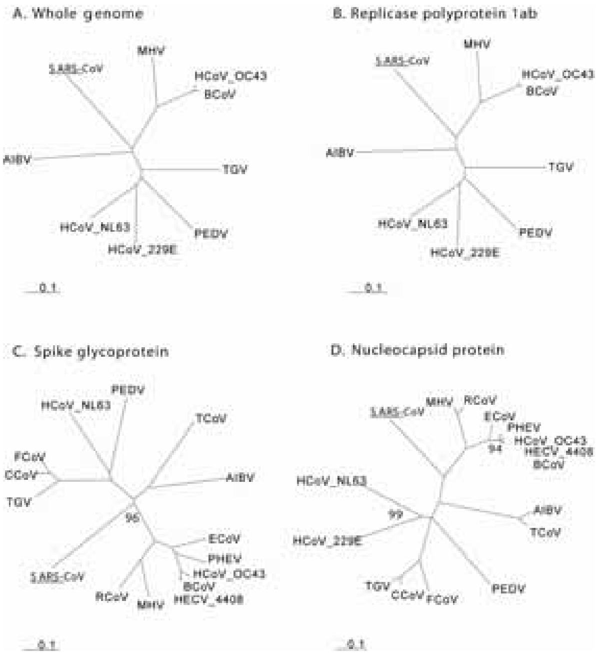
Ⅲ. Results & Discussion

In the result of the phylogenetic analysis using SARS-CoV genes with other coronavirus species, polyprotein 1ab and spike glycoprotein genes as well as the whole genome sequences, SARS-CoV showed distinct location from other coronavirus species including human-host coronaviruses such as HCoV-OC43, and HCoV-NL63 (Table 1, Fig. 1). AIBV also revealed distinct correlations with other coronaviruses, but it was not grouped with SARS-CoV, either. There was no specific correlation with SARS_CoV and FCoV whose reservoir was feline species such as cats and civets in this analysis. FCoV only showed close correlations with CCoV and TGV whose host reservoirs were dogs and pigs, respectively (Fig 1C, 1D).

Secondly, RSCU patterns were analyzed (Fig 2). SARS_CoV showed very low differences in RSCU values than those of other species, and FCoV followed those patterns very similarly (Fig. 2A). The overall RSCU patterns of SARS_CoV seemed similar with other coronaviruses showing the highest peaks in codon AGA for arginine, and GGU for glycine. In figure 2B, HIVs in *Lentivirus* revealed similar patterns with *simian immunodeficiency virus* (SIV) and *simian/human immunodeficiency virus* (SHIV), showing low RSCU differences, except for on the the codons which were encoded serine.

[Table 1] The list of the coronavirus species and their accession numbers which were used for the phylogenetic tree building in each gene.

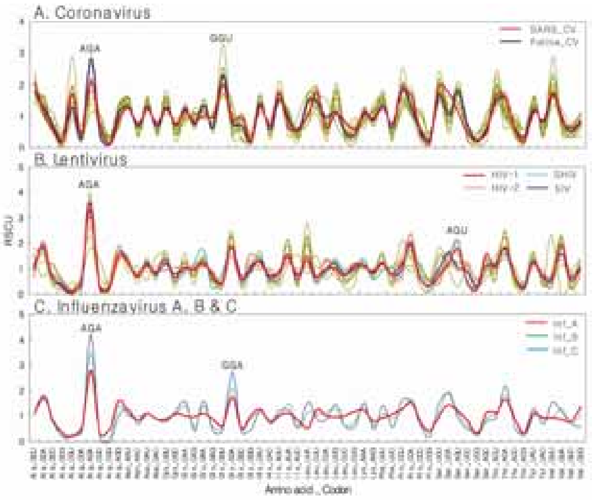
Species name	Abbreviation	Accession number
Avian infectious bronchitis virus	AIBV	NC_001451
Bovine coronavirus	BCoV	NC_003045
Canine coronavirus	CCoV	AY436637
Equine coronavirus	ECoV	AF251144
Feline coronavirus	FCoV	NC_007025
Human enteric coronavirus 4408	HECV-4480	AY316299
Human coronavirus 229E	HCoV-229E	AF304460
Human coronavirus NL63	HCoV-NL63	NC_005831
Human coronavirus OC43	HCoV-OC43	NC_005147
Murine hepatitis virus	MHV	NC_001846
Porcine epidemic diarrhea virus	PEDV	NC_003436
Porcine hemagglutinating encephalomyelitis virus	PHEV	DQ011855
Rat coronavirus	RCoV	AF088984
SARS coronavirus	SARS-CoV	NC_004718
Turkey coronavirus	TCoV	AY342357
Transmissible gastroenteritis virus	TGV	NC_002306



▶▶ Figure 1. Unrooted phylogenetic trees of the coronavirus species using (A) whole genome sequences, (B) replicase polyprotein 1ab gene, (C) spike glycoprotein gene, and (D) nucleocapsid protein gene. Bootstrap values (%) that were not 100 % are represented in each node.

Within this codon group, SHIV was seen more similar patterns to HIV-1 (red line), and SIV to HIV-2 (pink line). According to Reimann et al. [15], the envelopes of HIV-1 and SIV are quite genetically divergent, and SHIV is a recombinant chimeric SIV which can express the envelope genes of HIV-1. This might be one of the reasons why RSCU pattern in SHIV was more similar to HIV-1 than that of HIV-2. As for the three influenzavirus species, Inf_A showed the lowest differences in synonymous codon usage rates than other two influenzaviruses (Fig. 2C). Consequently, very low RSCU differences were shown commonly in all the target RNA viruses such as SARS_CoV, HIVs and Inf_A, and this patterns might provide some advantages for them to be transmitted from other species into human beings successfully.

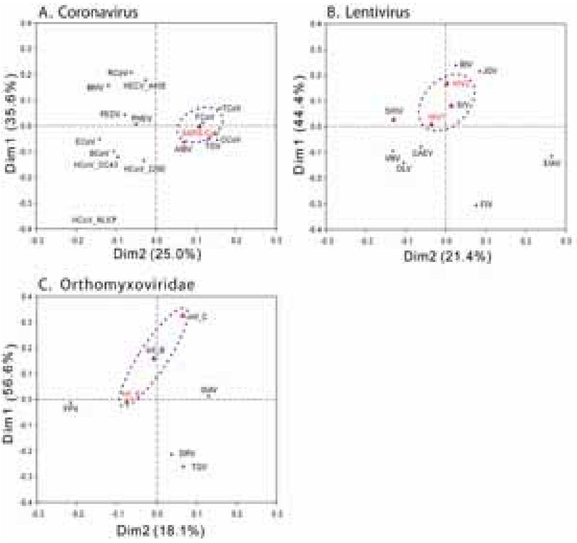
For statistical analysis for RSCU pattern differences among species, correspondence analysis (CA) using RSCU values of the species were performed (Fig. 3). Most of all, FCoV showed the closest correlation with SARS_CoV on the both axes of CA plots (Fig. 3A). This result was highly agreed with the serological study of Martina et al. [4], which reported that domestic cats and ferrets were susceptible to experimental infection by SARS_CoV, and that the virus was efficiently transmitted to animals living with them. Moreover, those who handled primarily civets were most likely to have antibodies, followed by traders who dealt mostly in wild boars and muntjac deers [2, 3].



▶▶ Figure 2. Relative synonymous codon usage (RSCU) values of viruses including (A) the genus *Coronavirus*, to which belongs SARS *coronavirus*, (B) the genus Lentivirus, to which belongs HIVs, and (C) the three genera, *Influenzavirus A, B* and *C* in 59 codons.

Although there was no phylogenetic correlation between SARS_CoV and FCoV (Fig. 1), CA result on the basis of RSCU values presented a strong clue for the close correlations between them. TGV and CCoV which were grouped together in phylogenetic analysis (Fig. 1C & 1D) were also located on the similar side in CA plots. SARS_CoV showed almost no bias on the first axis. As for the genus *Lentivirus* in figure 3B, SIV was located between HIV-1 and HIV-2 along with the first axis. SIV has been known as the origin of HIVs [16, 17, 18, 19], and CA plots showed this relationship very well. HIV-1 showed much less biased than HIV-2 on the basis of the first axis. Among the family *Orthomyxoviridae*, *influenza A,B, and C viruses* showed different RSCU patterns along with the first axis (Fig. 3C). As a result, these three target viruses including SARS_CoV, HIV-1, and Inf_A revealed much less biases than those of other related species on the first axis.

Synonymous codon usage bias was usually determined by the base compositions on the third codon position. According to the linear regression result, all the first axes showed significant correlations with the % GC content on each third codon, and it means that the first axis represented the RSCU bias in each virus species (data not shown). So, these target viruses seemed to have just a little or no bias when they use synonymous



▶▶ Figure 3. Plots of the values of the first and second axes of the genera (A) *Coronavirus*, (B) *Lentivirus*, and the family (C) *Orthomyxoviridae*. All the target pandemic viruses were presented in red colour.

codons during their infection stage.

In our analysis, RSCU values in SARS_CoV revealed very less biased than those of other coronaviruses, and FCoV followed that pattern very similarly. CA results in the genus *Coronavirus* also suggested that there was close relationship between SARS_CoV and FCoV. Until these days, SARS_CoV showed no specific correlations with other coronavirus species in phylogenetic analysis including this study, so some scientists started to call it as a fourth type of coronaviruses [5, 6, 20]. But, nonetheless, the close correlations between SARS_CoV and feline species such as domestic cats or civets were reported by many serological studies [2, 3, 4]. According to our results, synonymous codon usage patterns in all the CDSs of viruses might be able to detect serological correlations which could not be identified in phylogenetic analysis only.

Consequently, although there was no close correlation between SARS_CoV and other coronavirus species in phylogenetic analysis, it was possible to detect serological relationship with FCoV using synonymous codon pattern analysis. Moreover, all three target zoonotic viruses commonly showed very low RSCU bias than that of other related viruses. So, it might be helpful to use not only the phylogenetic analysis, but also the RSCU patterns analysis to identify the serologic characteristics of unknown pandemic virus species.

▌References▌

[1] Riley, S. et al., “Transmission Dynamics of the Etiological Agent of SARS in Hong Kong,” Science Vol. 300, pp. 1961-1966, 2003.

[2] Enserink, M., and Normile, D., "Search for SARS Origins Stalls," Science Vol. 302, pp. 766-767, 2003.

[3] Yu, D. et al., "Prevalence of IgG antibody to SARS-Associated Coronavirus in Animal Traders," CDC's Morbidity and Mortality Weekly Report Vol. 52, pp. 986-987, 2003.

[4] Martina, B.E.E. et al,. "SARS virus infection of cats and ferrets," Nature Vol. 425, p. 915, 2003.

[5] Gu, W. et al., "Analysis of synonymous codon usage in SARS Coronavirus and other viruses in the Nidovirales," Virus Res. Vol. 101, pp. 155-161, 2004.

[6] Peiris J.S.M. 2003. Severe Acute Respiratory Syndrome (SARS). J. Clin. Virol. 28:245-247.

[7] Peiris, J.S.M. et al., "Coronavirus as a possible cause of severe acute respiratory syndrome," Lancet Vol. 361, pp. 1322-1325, 2003.

[8] Rota, P.A. et al., "Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome," Science Vol. 300, pp. 1394-1399, 2003.

[9] Takebe, Y., Kusagawa, S., and Motomura, K., "Molecular epidemiology of HIV: Tracking AIDS pandemic," Pediatrics Int. Vol. 46, pp. 236-244, 2004.

[10] Worobey, M. et al. 2004. Contaminated polio vaccine theory refuted. Nature 428: 820.

[11] Thompson, J.D. et al., "CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools," Nucleic Acids Res. Vol. 25, pp. 4876-4882, 1997.

[12] Swofford, D.L., "PAUP*. Phylogenetic Analysis Using Parsimony, Version 4," Sinauer, Sunderland, MA, 1999.

[13] Sharp, P.M., and Li, W.H., "Codon usage in regulatory genes in *Escherichia coli* does not reflect for 'rare' codons," Nucleic Acids Res. Vol. 14, pp. 7737-7749, 1986.

[14] Cary, N.C., "SAS/IML User's Guide, Release 9.1 Edition," SAS Institute Inc., 1988.

[15] Reimann, K.A. et al., "An env gene derived from a primary human immunodeficiency virus type 1 isolate confers high in vivo replicative capacity to a chimeric simian/human immunodeficiency virus in rhesus monkey," J. Virol. Vol. 70, pp. 3198-3206, 1996.

[16] Dimmock, N.J., Easton, A.J., and Leppard, K.N., "Introduction to Modern Virology: fifth edition," Blackwell Publishing co., 2002.

[17] Gao, F. et al., "Origin of HIV-1 in the chimpanzee pan troglodytes," Nature Vol. 397, pp. 436-441, 1999.

[18] Hahn, B.H. et al., "AIDS as a zoonosis: scientific and public health implications," Science Vol. 287, pp. 607-614., 2000.

[19] Sharp, P.M. et al., "Origins and evolution of AIDS viruses: estimating the time scale," Biochem. Soc. Trans. Vol. 28, pp. 275-282, 2000.

[20] Marra, M.A. et al., "The genome sequence of the SARS-associated coronavirus," Science Vol. 300, pp. 1399-1404., 2003.