SARS-CoV Genome Polymorphism: A Bioinformatics Study

Gordana M. Pavlović-Lažetić[1]*, Nenad S. Mitić[1], Andrija M. Tomović[2], Mirjana D. Pavlović[3], and Miloš V. Beljanski[3]

[1] *Faculty of Mathematics, University of Belgrade, 11001 Belgrade, Serbia and Montenegro;* [2] *Friedrich Miescher Institute for Biomedical Research, CH-4058 Basel, Switzerland;* [3] *Institute of General and Physical Chemistry, 11001 Belgrade, Serbia and Montenegro.*

A dataset of 103 SARS-CoV isolates (101 human patients and 2 palm civets) was investigated on different aspects of genome **polymorphism and isolate classification. The number and the distribution of single nucleotide variations (SNVs) and insertions and deletions, with respect to a "profile", were determined and discussed ("profile" being a sequence containing the most represented letter per position). Distribution of substitution categories per codon positions, as well as synonymous and non-synonymous substitutions in coding regions of annotated isolates, was determined, along with amino acid (a.a.) property changes. Similar analysis was performed for the spike (S) protein in all the isolates (55 of them being predicted for the first time). The ratio Ka/Ks confirmed that the S gene was subjected to the Darwinian selection during virus transmission from animals to humans. Isolates from the dataset were classified according to genome polymorphism and genotypes. Genome polymorphism yields to two groups, one with a small number of SNVs and another with a large number of SNVs, with up to four subgroups with respect to insertions and deletions. We identified three basic nine-locus genotypes: TTTT/TTCGG, CGCC/TTCAT, and TGCC/TTCGT, with four subgenotypes. Both classifications proposed are in accordance with the new insights into possible epidemiological spread, both in space and time.**

Key words: SARS Coronavirus, single nucleotide polymorphism, insertions, deletions, spike protein, phylogenesis

Introduction

Severe acute respiratory syndrome (SARS), potentially fatal atypical pneumonia, first appeared in Guangdong province of China in November 2002 and soon afterward, within six months, spreaded all over the world (30 countries including China, Singapore, Vietnam, Canada, and USA), killing more than 700 people *(1)*. In less than four weeks after the global outbreak, a novel member of Coronaviridae family, namely SARS Coronavirus (SARS-CoV), was identified in the blood of respiratory specimens and stools of SARS patients, and confirmed as the causative agent of disease according to the Koch postulates *(2)*. Soon afterwards, first fully sequenced genomes of viral isolates were published *(3,4)*. In 2005 the number of fully sequenced viral isolates exceeds one hundred (http://www.ncbi.nlm.nih.gov/entrez).

SARS-CoV probably originated due to genetic exchange (recombination) and/or mutations between viruses with different host specificities (5, *6)*. Since coronaviruses are known to relatively easily jump among species, it was hypothesized that the new virus might have originated from wild animals. The analysis of SARS-CoV proteins supports and suggests possible past recombination event between mammalian-like and avian-like parent viruses *(6)*. Common sequence variants define three distinct genotypes of the SARS-CoV: one linked with animal [palm civet *(Paguma larvata)]* SARS-like viruses and early human phase, the other two linked with middle and late human phases, respectively *(7,8)*. SARS-CoV has a deleterious mutation of 29 nucleotides relative to the palm civet virus, indicating that if there was direct transmission, it went from civet to human, because deletions occur probably more easily than insertions (5). However, more recent reports indicate

* **Corresponding author.**
E-mail: gordana@matf.bg.ac.yu

that SARS-CoV is distinct from the civet virus and it has not been answered so far whether the SARS-CoV originated from civet, or civet was infected from other species *(9,10)*. The genome is relatively stable, since its mutation rate has been determined to be between 1.83×10^{-6} and 8.26×10^{-6} nucleotide substitutions per site per day *(11)*.

The SARS-CoV genome is approximately 30 Kb positive single strand RNA that corresponds to polycistronic mRNA, consisting of 5' and 3' untranslated regions (UTRs), 13 to 15 open reading frames (ORFs), and about 10 intergenic regions (IGRs) *(9,12,13)*. Its genome includes genes encoding two replicate polyproteins (RNA-dependant-RNA-polymerase, i.e., pp 1a and pp 1ab), encompassing two-thirds of the genome, and a set of ORFs at 3' end that code for four structural proteins: surface spike (S) glycoprotein (1,256 a.a.), envelope (E, 77 a.a.), matrix (M, 222 a.a.), and nucleocapsid (N, 423 a.a.) proteins. It also encodes for additional 8-9 predicted ORFs whose protein product functions are still under investigation *(14;* http://www.ncbi.nlm.nih.gov/entrez).

The S protein is the main surface antigen of the SARS-CoV and is involved in virus attachment on susceptible cells using mechanism similar to those of class I fusion proteins. The receptor for the SARS-CoV S protein is identified as angiotensin-converting enzyme 2 (ACE-2), which is a metallopeptidase *(15)*. The receptor-binding domain (RBD) has been determined to lay between a.a. postions 270-625 in recent studies *(16-20)*.

Several epitope sites, defined by polyclonal or monoclonal antibodies, have been identified on the S protein, depending on experimental conditions, all lying within wide or narrow regions between a.a. 12-1,192 *(20-31)*. Defining conserved immunodominant epitope regions of the S protein is of crucial importance for future anti-SARS vaccine development.

The main goal of this work was twofold: to perform mutation analysis of SARS-CoV viral genomes, with special attention to the S protein; and to group them according to different aspects of sequence similarity, eventually pointing to phylogeny and epidemiological dynamics of SARS-CoV.

Results and Discussion

Nucleotide content

Nucleotide content of SARS-CoV isolates favors T and A nucleotides. The corresponding percentages of letters in non-UTR regions of all the 96 isolates were found to be as follows: T (30.7940%), A (28.4246%), G (20.8121%), C (19.9535%), N (G, A, T, C; 0.0143%), R (Pur; 0.0005%), K (G or T; 0.0001%), M (A or C; 0.0002%), S (G or C; 0.0001%), W (A or T; 0.0002%), and Y (Pyr; 0.0004%). The overall ratio of (A,T)/(G,C) in the dataset was almost 3:2 (1.45). The ratio of Pur vs. Pyr nucleotides was almost 1 (0.97).

The distribution of nucleotides (nt) over sequences of length 250 nt is given in Figure S1 (Supporting Online Material). It exhibits three peak-regions of T nucleotide in the second quarter of the genome (ORF 1a), and rather stable behavior in the third quarter of the genome (ORF 1b), as also observed by Pyrc *et al (32)* for a group of coronaviruses (HCoV-NL63, HCoV-229E, SARS-CoV, and HCoV-OC43). Deviation of percentage of nucleotides over 250-nt blocks from the corresponding percentage in the whole dataset is given in Figure S2. Except for 3' UTR where T nucleotide is underrepresented with even about —13%, the highest excess from the average is about +10% in four peaks, which is exhibited again by T nucleotide, three of them being between positions 7,000 and 11,000 (ORF 1a), complementary with the nucleotide A represented with —10%, and the fourth one in the S protein. Otherwise the nucleotides' offset oscillates rather regularly between —5% and +5% from the average.

Genome polymorphism

All the isolates had high degree of nucleotide identity (more than 99% pair wise). Still, they could be differentiated on the basis of their genome polymorphism, i.e., the number and sites of SNVs and insertions and deletions (INDELs). Analysis of genomic polymorphism of the isolates resulted in the following two facts (Tables 1, S1, and S2). Firstly, two isolates, HSR 1 and AS, coincided with the "profile" on all the "non-empty" positions (see Materials and Methods) up to the poly-A sequence. Secondly, three isolates had large number of undefined nucleotides (N), either as contiguous segments (Sin3408 in ORFs 8a, 8b; Sin3408L in ORF 1b), or as scattered individual nucleotides or short clusters (SinP2) (Table S2). Isolate Sin3408 was the only one that has a 34-nt longer 5' UTR as compared with the "profile". Thus these three isolates were not considered to be reliably compared with others.