BIOINFORMATICS ANALYSUS OF SARS CoV S PROTEIN

S. Ranisavljević[1], S. Tadić[1], N. Tošić[1]
G. Pavlović-Lažetić[1], N. Mitić[1],M. D. Pavlović[2] and S. Ž. Gorjanović[2]
[1]*Faculty of Mathematics, University of Belgrade, Belgrade, Serbia*
[2]*Institute of General and Physical Chemistry, Belgrade, Serbia*

ABSTRACT
Bioinformatic study of SARS-CoV S protein for 148 fully sequenced isolates was performed, including different aspects of genome polymorphism (i.e. single nucleotide polymorphisms - SNP), amino acids properties changes.

INTRODUCTION

Severe acute respiratory syndrome (SARS) is a respiratory disease in humans caused by the SARS coronavirus (CoV). The SARS CoV genome is approximately 30 Kb positive single strand RNA that corresponds to polycistronic mRNA, consisting of 5- and 3- untranslated regions, up to 15 open reading frames (ORFs) and about 10 intergenic regions. Spike (S) protein is of special interest because of its role in viral attachment and host range determination. It is coded by ORF 2 and consists of 1255 amino acids (aa), with at least three domains: external (aa 1-1196), that includes signal peptide (aa 1-14), receptor binding domain (RBD, aa 270-625) and several experimentally determined epitope sites, transmembrane (aa 1196-1218), and internal (aa1218-1255) domains (Fig. 1. B.) [1, 2].

In this work we performed bioinformatics study (i.e. single nucleotide polymorphisms - SNPs, amino acids properties changes etc.) of SARS-CoV S protein for 148 fully sequenced isolates.

EXPERIMENTAL

Nucleotide sequences of 148 SARS-CoVs and S proteins were taken from PubMed NCBI Entrez [3] database in GenBank and Fasta format. We located start and end of S protein in sequences in which it has not been annotated with ORF finder tool [4]. With use of Extracseq and Transseq programs from EMBOSS [5] package we extracted S protein nucleotide sequences and translated them to aa sequences. All 148 nucleotide and aa sequences have been aligned with ClustalW program [6]. Statistic was generated using Biom [7], software which we originally developed using Java programming language, which determines all statistics represented in this paper.

RESULTS AND DISCUSSION

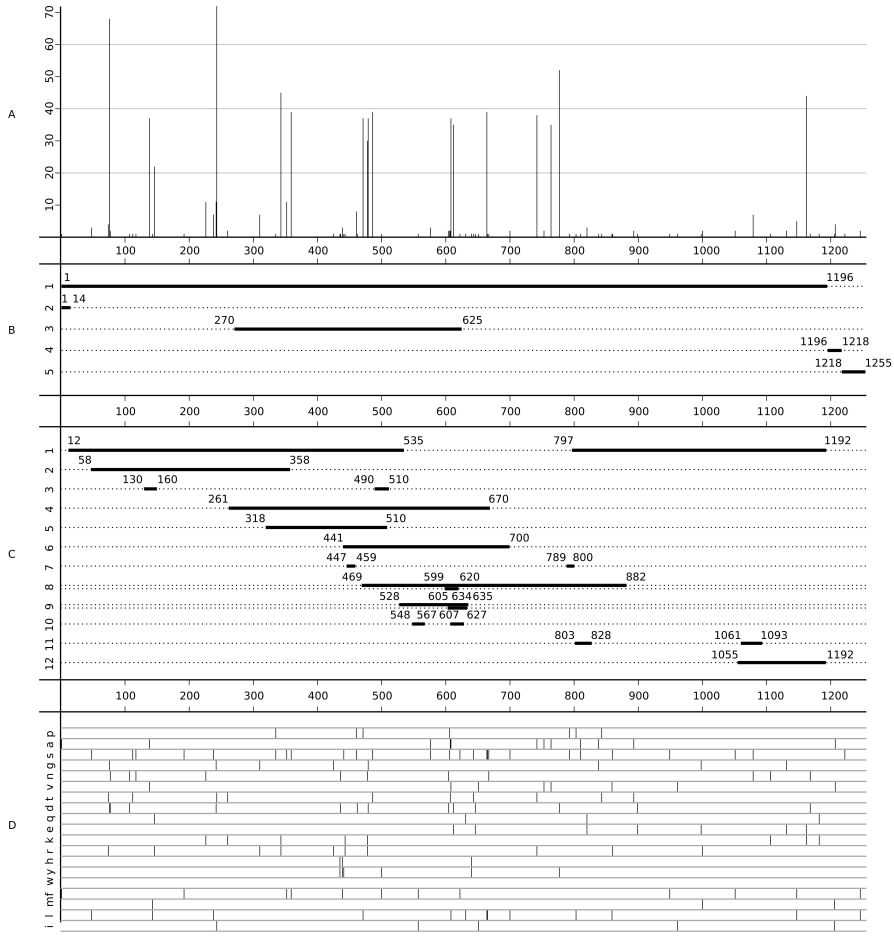

Figure 1.

Position of aa changes and their number relatively to average "profile" isolate S protein (for all of 148 SARS CoV isolates) are shown in Fig. 1. A. S protein domains (Fig 1. B) and experimentally determinated epitopes from ref. [1] are shown on Fig. 1. C, while positions of all determinate aa changes relatively to each other are revealed at Fig 1. D.

Coding sequence analysis. In all of 148 S protein gene sequences, there is 119 SNPs positions. There is totally 1146 SNPs, 856 and 290 of which are transitions and transversions, respectively. Most common nucleotide substitution is T→C which happened 284 times favoured by the preceding T (125 times) and the followed by T (151 times). A→G nucleotide substitution happened 271 times

favoured by the preceding A and following A (124 times and 145 times, respectively). Nucleotide substitution A→T happened only 2 times favoured by the preceding T (one time) and following A (one time). Nucleotide substitution C→A happened 6 times favoured by the preceding A and following A (4 times and 5 times, respectively). SNPs are found on first, second and third place in codon 36, 44 and 39 times respectively. Consequently, number of silent mutations (without aa changes) is 32, number of mutations without aa properties changes is 2, and there are 85 mutations with amino acid properties changes.

Protein sequence analysis. Changes in S protein external domain are of specific importance, since it correlates with altered pathogenesis, virulence and may be important for vaccine design [1, 2]. Total number of aa substitutions in external domain is 83, where 82 substitutions have aa properties changed in respect to profile sequence and may led to different structural forms of S protein. There are 26 aa substitutions in RBD (all having change of aa properties) in 65 isolates. These changes may lead to wider range of cellular receptors for viral binding. Since experimentally determined epitopes overlap almost whole external domain, (including RBD), all of these changes are significant for future vaccine development [1, 2].

CONCLUSION

There are 83 aa changes in 148 fully (up to now) sequenced isolates. Most of them (82) have changed aa properties that may lead to wider range of cellular receptors for viral binding and also may be important in production of protective agents against SARS CoV.

ACKNOWELAGEMENT: This study is undergraduate students exercise in Bioinformatics, under guidance of Dr M. V. Beljanski and is supported by Ministry of Science, Republic of Serbia, Project No. 144030.

REFERENCES:

[1] G. M. Pavlović-Lažetić, N. S. Mitić, A. M. Tomović, M. D. Pavlović, M. V. Beljanski, Geno. Prot. Bioinfo., 2005, **3**, 18-34.
[2] C. R. Astell, R. A. Holt, S. J. M. Jones, M. A. Marra, in: Coronaviruses with Special Emphasis on First Insights Concerning SARS, A. Schmidt, M.H. Wolff and O. Weber (Eds.), Birkhäuser Verlag, Basel, 2005
[3] http://www.ncbi.nlm.nih.gov/entrez
[4] http://www.ncbi.nlm.nih.gov/gorf/gorf.html
[5] http://emboss.sourceforge.net/
[6] http://www.ebi.ac.uk/clustalw
[7] http://biom.sourceforge.net/