Research Article

GENOMIC AND PROTEOMIC STUDIES USING COMPUTATIONAL APPROACHES IN SARS GENOME

Dr. DSVGK Kaladhar
Department of Bioinformatics, GITAM University, Visakhapatnam.
***Corresponding author's E-mail:** dr.dowluru@gmail.com

Accepted on: 30-01-2011; **Finalized on:** 10-04-2011.

ABSTRACT

Integration of Biological studies and Information technology in research studies made to identify most of the hidden secrets of life to the society. The present studies provided genomic and proteomic studies of SARS virus in application of computational tools. Genomic sequence of SARS virus (29,751 bp) from NCBI database predicted eleven potential genes. The sequences shown relationship with Porcine hemagglutinating encephalomyelitis virus and RNA murine hepatitis virus. The results also predicts that acyltransferase family proteins of Lactobacillus acidophilus and nucleocapsid of SARS coronavirus are extremely basic in nature and spike glycoprotein of SARS coronavirus predicted acidic in nature.

Keywords: SARS virus, Genomics, Proteomics.

INTRODUCTION

Genomics, Proteomics and Metabolomics are new, emergent areas of science that uses computational approaches to solve biological problems[1]. The investigators take advantage of large, complex data sets in a vigorous fashion to reach valid, biological conclusions[2]. In the genome age, a major research goal is to find the functions of genes obtained from genomes and to define their interactions in a particular organism obtained by experiment[3]. Bioinformatics and computational biology involve the use of techniques to solve biological problems usually on the molecular level[4], such as drug discovery, prediction of molecular function and medical diagnosis[5].

The first epidemic of severe acute respiratory syndrome (SARS) started in Guangdong Province, China, spread by close person-to-person contact[6, 7, 8]. Various biological researchers has characterized the SARS virus and determining how to control it. Scientific information provided the society how the humans will respond and communicate in the setting of the next pandemic[9].

Coronaviruses characterized into three groups based on infection; groups 1 and 2 contain mammalian viruses, and group 3 contains avian viruses [10]. Each group of coronaviruses are again classified into distinct species based on host range, capsid architecture, antigenic relationships, and genomic organization. The size of SARS-CoV genome is of 29,727-nucleotide, polyadenylated RNA, and 41% of the residues contains G or C [11].

The genome sequence of SARS virus will aid in the diagnosis of SARS infection in humans and potential animal hosts (using polymerase chain reaction and immunological tests), in the development of antiviral compounds (including neutralizing antibodies), and in the identification of putative epitopes for vaccine development [12].

MATERIALS AND METHODS

The NCBI's (National Centre for Biotechnology research) database is a popular starting point for identifying nucleotides and genomes of different species. A complete genome sequence of selected SARS Virus Accession number NC_004718 is retrieved from the NCBI databank and is analyzed for the present study.

FGENESV0 algorithm is based on pattern recognition of different types of signals and Markov chain models of coding regions. Complete sequence of SARS genome has been submitted for FGENESV0 for the prediction of codons and genes.

Backtranseq command from EMBOSS accepts a protein sequence as input and uses a codon usage table to generate a DNA sequence representing the most likely non-degenerate coding sequence. A consensus sequence derived from all the possible codons for each amino acid is also shown based on regions of promoter and reading frames.

BLASTP takes protein (amino acid) sequences and compares them against the NCBI protein databases. This search is similar to the standard protein-protein BLAST with the parameters set to optimize for searching with short sequences. Because of its design for speed, there may be a minimal loss of sensitivity to distant sequence relationships.

BLASTN takes nucleotides sequences and compares them against the NCBI nucleotide databases. It is better at finding sequences similar, but not identical, to the query sequence.

Protein Molecular Weight accepts a protein sequence and calculates the molecular weight. Protein Isoelectric Point calculates the theoretical pI (isoelectric point) for the protein sequence entered (ExPASy web server - http://expasy.org/tools/pi_tool.html)

SWISS-MODEL is an automated protein structure homology-modeling server, which can be accessible via the ExPASy web server. The purpose of Swiss-model server is to make Protein modeling, accessible to all biochemists and molecular biologists World Wide.

RESULTS AND DISCUSSION

A complete genome sequence of SARS virus (ACCESSION NC_004718) is retrieved from the NCBI databank. The genomic sequence is taken as FASTA format from NCBI's Genbank flat file NCBI and is submitted to FGENESV0 server. Figure 1 shown result of 11 predicted potential genes (predicted proteins). Protein 1 (265-13413) contains largest amino acid sequence (4382 amino acids) and protein 7(26117-26347) predicted shortest aminoacid sequence (76 amino acids) from 29,751 bp of genome sequence.

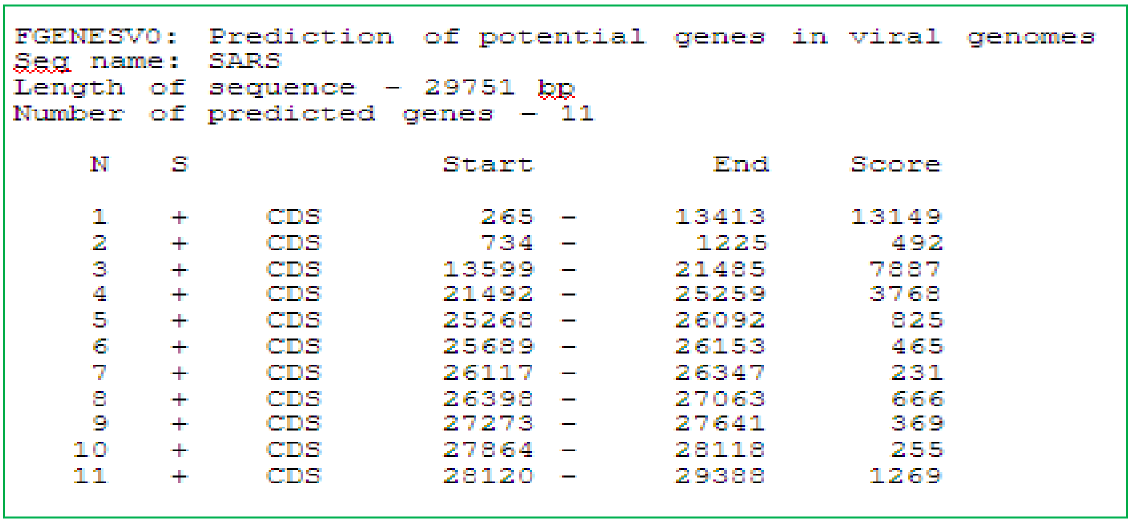
The reverse translation of protein sequence is done using EMBOSS. The reverse translated sequence and predicted proteins were submitted to BLAST. The BLASTN reports of reverse translated nucleic acid showed relationship with SARS complete genome from various isolates. BLASTP results of predicted proteins showed some relationship with other viruses such as Porcine hemagglutinating encephalomyelitis virus and murine hepatitis virus.

Protein1 predicted as putative polyprotein of SARS coronavirus showing relationship with replicase polyprotein of Porcine hemagglutinating encephalomyelitis virus and RNA-directed RNA polymerase of murine hepatitis virus. **Protein2** predicted as uncharacterized protein 1c of SARS coronavirus and acyltransferase family protein of Lactobacillus acidophilus. **Protein3** predicted as putative polyprotein of SARS coronavirus showing relationship with replicative polyprotein of Murine hepatitis virus. **Protein4** predicted as spike glycoprotein of SARS coronavirus showing relationship with spike glycoprotein precursor of murine hepatitis virus. **Protein5** predicted as hypothetical protein of SARS coronavirus showing relationship with hypothetical protein of Magnaporthe grisea. **Protein6** predicted as putative uncharacterized protein of SARS coronavirus showing relationship with far-red impaired response protein-like protein of Oryza sativa(japonica cultivar-group) and myosin IXB isoform 4 of Pan troglodytes. **Protein7** predicted as envelope protein of Bat SARS coronavirus showing relationship with small membrane protein of Human coronavirus. **Protein8** predicted as matrix protein of SARS coronavirus showing relationship with membrane protein of Porcine hemagglutinating encephalomyelitis virus. **Protein9** predicted as hypothetical protein of SARS coronavirus showing relationship with protein of Xenopus laevis. **Protein10** predicted as hypothetical protein of SARS coronavirus showing relationship with Hypothetical protein of Caenorhabditis briggsae. **Protein11** predicted as nucleocapsid protein of SARS coronavirus showing relationship with nucleocapsid protein of Murine hepatitis virus (Table 1).

The molecular weight and Isoelectric points of 11 proteins predicted from SARS genome are shown in Table 2. The following results predicts that acyltransferase family proteins of Lactobacillus acidophilus and nucleocapsid of SARS coronavirus are extremely basic in nature and spike glycoprotein of SARS coronavirus shows acidic in nature.

Predicted proteins are submitted to SwissModel for 3D structural analysis (Figure 2). Templates were not predicted in proteins 2, 5, 6, 7, 8 and 10. The protein models (swissModel) were provided for other proteins along with structure opening in Rasmol software. Highest numbers of helix structures were found in protein 1, highest numbers of strands and H-bonds were found in protein 3.

FIGURE 1: PREDICTION OF POTENTIAL GENES IN SARS VIRAL GENOME USING FGENESV0



```
FGENESV0:  Prediction  of  potential  genes  in  viral  genomes
Seq  name:  SARS
Length  of  sequence  -  29751  bp
Number  of  predicted  genes  -  11

     N    S              Start         End     Score

     1    +    CDS         265  -    13413     13149
     2    +    CDS         734  -     1225       492
     3    +    CDS       13599  -    21485      7887
     4    +    CDS       21492  -    25259      3768
     5    +    CDS       25268  -    26092       825
     6    +    CDS       25689  -    26153       465
     7    +    CDS       26117  -    26347       231
     8    +    CDS       26398  -    27063       666
     9    +    CDS       27273  -    27641       369
    10    +    CDS       27864  -    28118       255
    11    +    CDS       28120  -    29388      1269
```

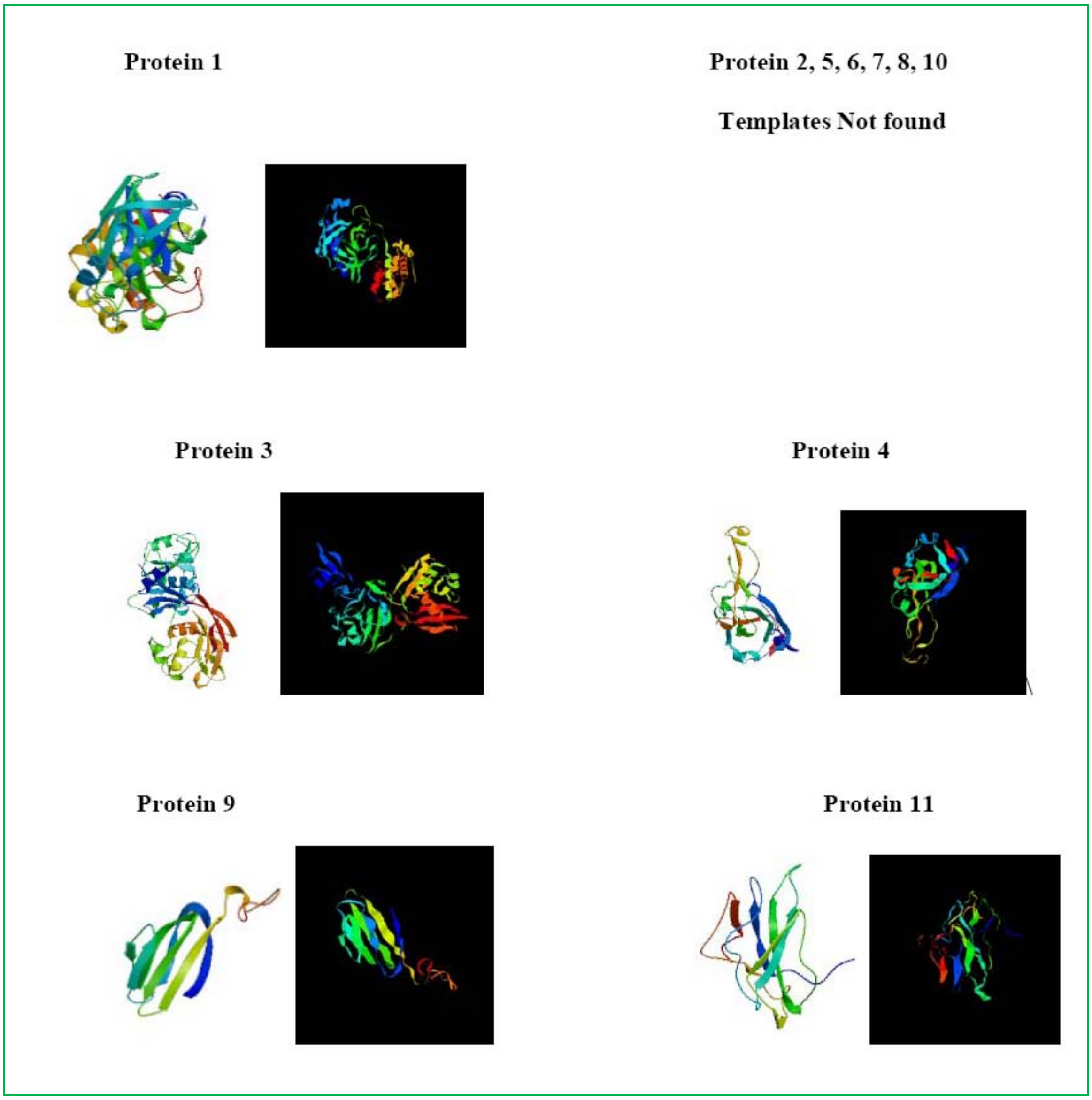
TABLE 1: RESULTS OF BLASTN AND BLASTP IN DECREASING ORDER OF SCORE

NUCLIC ACID / PROTEIN NUMBER	NUCLIC ACID (BLASTN)	PROTEIN (BLASTP)
1	• SARS coronavirus TW10, complete genome • SARS coronavirus TWY genomic RNA, complete genome • SARS coronavirus TW8, complete genome	• putative orf1ab polyprotein [SARS coronavirus PUMC01] • putative polyprotein [SARS coronavirus TW8]. • replicase polyprotein [Porcine hemagglutinating encephalomyelitis virus]. • RNA-directed RNA polymerase [murine hepatitis virus].
2	• SARS coronavirus GD01, complete genome • SARS coronavirus WHU, complete genome • SARS coronavirus TW3, complete genome	• uncharacterized protein 1c [SARS coronavirus ZJ0301] • acyltransferase family protein [Lactobacillus acidophilus NCFM]
3	• SARS coronavirus HC/SZ/61/03, complete genome • SARS Coronavirus CDC#200301157, complete genome • SARS coronavirus Sin847, complete genome	• putative polyprotein [SARS coronavirus TW9] • putative polyprotein [SARS coronavirus TW4]. • replicative polyprotein 1ab [Murine hepatitis virus].
4	• SARS coronavirus TOR2, complete genome • SARS coronavirus TWS genomic RNA, complete genome • SARS coronavirus isolate CUHKtc55NS spike glycoprotein (S)	• spike glycoprotein [SARS coronavirus]. • Spike glycoprotein precursor (S glycoprotein) (Peplomer protein)(E2) [Contains: Spike protein S1; Spike protein S2] SARS coronavirus. • spike glycoprotein precursor [Murine hepatitis virus].
5	• SARS coronavirus LLJ-2004, complete genome • SARS coronavirus BJ02, complete genome • SARS coronavirus BJ202, complete genome	• hypothetical protein sars3a [SARS coronavirus] • 3a [SARS coronavirus LLJ-2004]. • hypothetical protein MGG_12698 [Magnaporthe grisea 70-15].
6	• SARS coronavirus GDH-BJH01, complete genome • SARS coronavirus BJ202, complete genome • SARS coronavirus BJ162, complete genome	• putative uncharacterized protein 2 [SARS coronavirus BJ01]. • 3b [SARS coronavirus LLJ-2004]. • far-red impaired response protein-like protein [Oryza sativa(japonica cultivar-group)]. • myosin IXB isoform 4 [Pan troglodytes].
7	• SARS coronavirus GDH-BJH01, complete genome • SARS coronavirus BJ202, complete genome • SARS coronavirus BJ162, complete genome	• envelope protein [Bat SARS coronavirus Rm1] • envelope protein [SARS coronavirus Sino3-11]. • small membrane protein [Human coronavirus HKU1].
8	• SARS coronavirus GD322 M protein gene, complete cds • SARS coronavirus GZ02, complete genome • SARS coronavirus ZS-C, complete genome	• matrix protein [SARS coronavirus] • M protein [SARS coronavirus Urbani]. • membrane protein [Porcine hemagglutinating encephalomyelitis virus].
9	• SARS coronavirus isolate HC/SZ/266/03, complete genome • Bat coronavirus (BtCoV/279/2005), complete genome • Bat SARS coronavirus Rm1, complete genome	• hypothetical protein sars7a [SARS coronavirus]. • 7a [SARS coronavirus LLJ-2004]. • protein [Xenopus laevis]. (African clawed frog)
10	• SARS coronavirus HKU-65806, partial genome • SARS coronavirus GDH-BJH01, complete genome • SARS coronavirus BJ202, complete genome.	• hypothetical protein sars8b [SARS coronavirus]. • 8 [SARS coronavirus LLJ-2004]. • Hypothetical protein CBG05593 [Caenorhabditis briggsae].
11	• SARS coronavirus GZ02, complete genome • SARS coronavirus HGZ8L1-A, partial genome • SARS coronavirus TJ01 nucleocapsid protein gene.	• nucleocapsid protein [SARS coronavirus]. • putative nucleocapsid protein N [SARS coronavirus TW11]. • nucleocapsid protein [Murine hepatitis virus].

TABLE 2: PROTEIN MOLECULAR WEIGHT AND ISOELECTRIC POINT

PROTEIN NUMBER	MOLECULAR WEIGHT (in kilodaltons)	ISOELECTRIC POINT (pH)
1	486.44	6.25
2	17.74	11.98
3	296.8	7.03
4	139.13	5.60
5	31.01	6.38
6	17.75	11.50
7	8.36	6.30
8	25.06	9.90
9	13.94	8.12
10	9.56	9.43
11	46.04	10.74

FIGURE 2: MODELING OF PROTEINS USING SWISSMODEL (STRUCTURES FROM SWISSMODEL AND RASMOL)



CONCLUSION

The coronaviruses (order Nidovirales, family Coronaviridae, genus Coronavirus) are a diverse group of large, enveloped, positive-stranded RNA viruses that cause respiratory and enteric diseases in humans and other animals[13]. Most of the biologists focus to explore innovations of their research in faster rate using developments in Information technology, using bioinformatics tools[14]. Biological research provides deeper insights into the complexity of living organisms while computer science provides an effective means to store and analyze large volumes of complex data[15].

Acknowledgement: The author would like to thank GITAM University for providing computational and wet lab facility and access to e-journals to carry out this research.

REFERENCES

1. Andrea DW and Leroy H, Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine, Journal of Proteome Research, 3 *(2),* 2004**,** 179-196.
2. Andreas DB and. Francis BFO, Bioinformatics- A practical guide to the analysis of genes and proteins, A John Wiley & Sons, Inc. Publication, Second edition, 2001, 1.
3. David WM, Bioinformatics sequence and genome analysis second edition, CBS publishers, 2005, 3-7.
4. Yi-Ping PC, Feng C, Identifying targets for drug discovery using bioinformatics, Expert Opinion on Therapeutic Targets, 12(4), 2008, 383-389.
5. Nageswara PVR, Uma TD, Kaladhar DSVGK, Sridhar GR, Allam AR. A probabilistic neural network approach for protein superfamily classification, JATIT, 6(1), 2009, 101-105.
6. Peiris JSM, Lai ST, Poon LLM, Guan Y, Yam LYC, Lim W, Nicholls J, Yee WKS, Yan WW, Cheung MT, Cheng VCC, Chan KH, Tsang DNC, Yung RWH, Ng TK, Yuen KY, Coronavirus as a possible cause of severe acute respiratory syndrome, Lancet, 361(9366)**,** 2003, 1319-1325.
7. Kathryn VH, SARS-Associated Coronavirus, N Engl J Med, 348, 2003, 1948-1951.
8. Leo LMP, Cynthia SWL, Masato T, Kwok HC, Bonnie WYW, Kwok YY, Yi G, Joseph SMP, Rapid Detection of the Severe Acute Respiratory Syndrome (SARS) Coronavirus by a Loop-Mediated Isothermal Amplification Assay, Clinical Chemistry, 50, 2004, 1050-1052.
9. Gronvall GK, Waldhorn RE, Henderson DA, The Scientific Response to a Pandemic, PLoS Pathogens, 2(2), 2006, e9.
10. Teklu K, Daniel A, Update on virological, epidemiological and diagnostic aspects of Sars-Corona Virus (SARS-CoV): A newly emerging virus, Ethiop.J.Health Dev, 18(1), 2004, 52-54.
11. Paul AR, Steven MO, Stephan SM, Allan WN, Ray C, Joseph PI, Silvia P, Bettina B, Kaija M, Min-hsin C, Suxiong T, Azaibi T, Luis L, Michael F, Joseph LDR, Qi C, David W, Dean DE, Teresa CTP, Cara B, Thomas GK, Pierre ER, Anthony S, Stephanie L, Brian H, Josef L, Karen MC, Melissa OR, Ron F, Stephan G, Albert DMEO, Christian D, Mark AP, Larry JA, William JB, Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome, Science , 300(5624), 2003, 1394-1399.
12. Marco AM, Steven JMJ, Caroline RA, Robert AH, Angela BW, Yaron SNB, Jaswinder K, Jennifer KA, Sarah AB, Susanna YC, Alison C, Shaun MC, Doug F, Noreen G, Obi LG, Stephen RL, Michael M, Helen MD, Stephen BM, Pawan KP, Anca SP, Gordon AR, Jacqueline ES, Asim S, Duane ES, Jeff MS, George SY, Francis P, Anton A, Harvey A, Nathalie B, Kathy B, Timothy FB, Donnie B, Martin C, Michael D, Lisa F, Ramon F, Michael G, Michael G, Allen G, Steven J, Heinz F, Adrienne M, Amin K, Yan L, Susan N, Ute S, Graham AT, Shaun T, Robert V, Diane W, Brynn W, Robert CB, Mel K, Martin P, Danuta MS, Chris U, Rachel LR, The Genome Sequence of the SARS-Associated Coronavirus, Science ,300(5624), 2003, 1399-1404.
13. Stanley P, Jason N, Coronaviruses post-SARS: Update on replication and pathogenesis, Nat Rev Microbiol., 7(6), 2009, 439-450.
14. Kaladhar DSVGK, Uma DT, Nageswara RPV, An in silico genome wide identification, characterization and modeling of Human papilloma virus strain 92, IJEST, 2(9), 2010, 4288-4291.
15. Kaladhar DSVGK, Jaya KB, Krishna CA, Molecular modeling and immunoinformatics studies in Swine Influenza H1N1 Genome, Journal of Pharmacy Research, 3(12), 2010, 3151-3154.

****