



# Genome-wide analysis of 10664 SARS-CoV-2 genomes to identify virus strains in 73 countries based on single nucleotide polymorphism

Nimisha Ghosh<sup>a,1</sup>, Indrajit Saha<sup>b,1,\*</sup>, Nikhil Sharma<sup>c,1</sup>, Suman Nandi<sup>b</sup>, Dariusz Plewczynski<sup>d,e</sup>

<sup>a</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

<sup>b</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

<sup>c</sup> Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

<sup>d</sup> Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland

<sup>e</sup> Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Warsaw, Poland

## ARTICLE INFO

### Keywords:

Clustering  
COVID-19  
Multiple sequence alignment  
Non-synonymous SNP  
SARS-CoV-2

## ABSTRACT

Since the onslaught of SARS-CoV-2, the research community has been searching for a vaccine to fight against this virus. However, during this period, the virus has mutated to adapt to the different environmental conditions in the world and made the task of vaccine design more challenging. In this situation, the identification of virus strains is very much timely and important task. We have performed genome-wide analysis of 10664 SARS-CoV-2 genomes of 73 countries to identify and prepare a Single Nucleotide Polymorphism (SNP) dataset of SARS-CoV-2. Thereafter, with the use of this SNP data, the advantage of hierarchical clustering is taken care of in such a way so that Average Linkage and Complete Linkage with Jaccard and Hamming distance functions are applied separately in order to identify the virus strains as clusters present in the SNP data. In this regard, the consensus of both the clustering results are also considered while Silhouette index is used as a cluster validity index to measure the goodness of the clusters as well to determine the number of clusters or virus strains. As a result, we have identified five major clusters or virus strains present worldwide. Apart from quantitative measures, these clusters are also visualized using Visual Assessment of Tendency (VAT) plot. The evolution of these clusters are also shown. Furthermore, top 10 signature SNPs are identified in each cluster and the non-synonymous signature SNPs are visualised in the respective protein structures. Also, the sequence and structural homology-based prediction along with the protein structural stability of these non-synonymous signature SNPs are reported in order to judge the characteristics of the identified clusters. As a consequence, T85I, Q57H and R203M in NSP2, ORF3a and Nucleocapsid respectively are found to be responsible for Cluster 1 as they are damaging and unstable non-synonymous signature SNPs. Similarly, F506L and S507C in Exon are responsible for both Clusters 3 and 4 while Clusters 2 and 5 do not exhibit such behaviour due to the absence of any non-synonymous signature SNPs. In addition to all these, the code, SNP dataset, 10664 labelled SARS-CoV-2 strains and additional results as supplementary are provided through our website for further use.

## 1. Introduction

SARS-CoV-2 is the causal agent for current ongoing outbreak of disease commonly known as COVID-19 (Zhou et al., 2020) which has proven to have a detrimental effect on the humankind. As a result, medical emergencies have surged and a halt of economic growth has occurred around the globe due to an eccentric impact of SARS-CoV-2.

SARS-CoV-2 belongs to the family of Coronaviridae which also houses SARS-CoV-1 and MERS-CoV (van Dorp et al., 2020). First case of Severe Acute Respiratory Syndrome (SARS) was registered way back in 2002-03, which took around 8000 lives.<sup>2</sup> Another pathogenic invasion was reported in 2012, named as Middle East Respiratory Syndrome Coronavirus (MERS-CoV) with a worldwide mortality rate of 35.5% (Ahmed, 2017). However, these two viruses have a significantly low

\* Corresponding author.

E-mail address: [indrajit@nitttrkol.ac.in](mailto:indrajit@nitttrkol.ac.in) (I. Saha).

<sup>1</sup> Equally contributed.

<sup>2</sup> [https://en.wikipedia.org/wiki/2002%E2%80%9303\\_SARS\\_outbreak](https://en.wikipedia.org/wiki/2002%E2%80%9303_SARS_outbreak).

<https://doi.org/10.1016/j.virusres.2021.198401>

Received 8 December 2020; Received in revised form 23 February 2021; Accepted 16 March 2021

Available online 26 March 2021

0168-1702/© 2021 Elsevier B.V. All rights reserved.

Nimisha Ghosh

Nimisha Ghosh



# A review on evolution of emerging SARS-CoV-2 variants based on spike glycoprotein

Nimisha Ghosh<sup>a,b,1</sup>, Suman Nandi<sup>c,1</sup>, Indrajit Saha<sup>c,1,\*</sup>

<sup>a</sup> Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland

<sup>b</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India

<sup>c</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

## ARTICLE INFO

### Keywords:

COVID-19

Mutations

SARS-CoV-2 genomes

Spike glycoprotein

Virus strains

## ABSTRACT

Since the inception of SARS-CoV-2 in December 2019, many variants have emerged over time. Some of these variants have resulted in transmissibility changes of the virus and may also have impact on diagnosis, therapeutics and even vaccines, thereby raising particular concerns in the scientific community. The variants which have mutations in Spike glycoprotein are the primary focus as it is the main target for neutralising antibodies. SARS-CoV-2 is known to infect human through Spike glycoprotein and uses receptor-binding domain (RBD) to bind to the ACE2 receptor in human. Thus, it is of utmost importance to study these variants and their corresponding mutations. Such 12 different important variants identified so far are B.1.1.7 (Alpha), B.1.351 (Beta), B.1.525 (Eta), B.1.427/B.1.429 (Epsilon), B.1.526 (Iota), B.1.617.1 (Kappa), B.1.617.2 (Delta), C.37 (Lambda), P.1 (Gamma), P.2 (Zeta), P.3 (Theta) and the recently discovered B.1.1.529 (Omicron). These variants have 84 unique mutations in Spike glycoprotein. To analyse such mutations, multiple sequence alignment of 77681 SARS-CoV-2 genomes of 98 countries over the period from January 2020 to July 2021 is performed followed by phylogenetic analysis. Also, characteristics of new emerging variants are elaborately discussed. The individual evolution of these mutation points and the respective variants are visualised and their characteristics are also reported. Moreover, to judge the characteristics of the non-synonymous mutation points (substitutions), their biological functions are evaluated by PolyPhen-2 while protein structural stability is evaluated using I-Mutant 2.0.

## 1. Introduction

The ongoing wave of COVID-19 caused by SARS-CoV-2 virus was first identified in the city of Wuhan, China during December 2019. Since then, the virus has spread very rapidly and has affected millions of people worldwide. SARS-CoV-2 is a positive stranded RNA virus with a length of about 30 kb encompassing non-structural and structural proteins. Spike glycoprotein, a structural protein present on the virus surface plays an important role in binding with ACE2. This RNA virus can make a replica of its own after binding with the host cell, thereby causing several mutations [24]. Whenever the mutation is significant, the structure of the virus changes, resulting in a new variant or lineage<sup>2</sup>

of the virus [38]. Motivated by this observation, in this study we have performed a competitive analysis of several variants of SARS-CoV-2. The mutation of SARS-CoV-2 is happening over time, thereby resulting in new variants. Whenever a new variant emerges, it can be called as an “emerging variant” which have some potential consequences viz. increase in transmissibility, morbidity as well as mortality. It is to be noted that the different variants have some unique as well as some common mutations. In this regard, there are 12 important variants as declared by W.H.O<sup>3</sup> and 84 unique mutations that are reported in this work. Some of these variants have been categorised as either variants of concern, variants of interest or variants under monitoring based on their transmissibility, immunity and infection severity<sup>4</sup>. As of now, the variants of

\* Corresponding author.

E-mail address: [indrajit@nittrkol.ac.in](mailto:indrajit@nittrkol.ac.in) (I. Saha).

<sup>1</sup> Equally contributed.

<sup>2</sup> [https://cov-lineages.org/lineages/lineage\\_B.1.html](https://cov-lineages.org/lineages/lineage_B.1.html).

<sup>3</sup> <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>.

<sup>4</sup> <https://www.ecdc.europa.eu/en/covid-19/variants-concern>.

Nimisha Ghosh



## Computers & Electrical Engineering

Volume 64, November 2017, Pages 288-304

# An efficient trajectory based routing scheme for delay-sensitive data in wireless sensor network ☆

Nimisha Ghosh  , Riddhiman Sett , Indrajit Banerjee 

Show more 

 Share  Cite

<https://doi.org/10.1016/j.compeleceng.2017.06.003>

[Get rights and content](#)

## Abstract

Exploiting sink mobility for balancing the energy dissipation of sensor nodes and preventing energy holes is a growing trend in the field of wireless sensor network (WSN). The mobile sink moves throughout the network and collects data in a single-hop fashion resulting in less energy consumption of the sensor nodes. Despite the advantage of this scheme, data forwarding to the mobile sink during an emergency situation is a major issue that should be dealt with. For efficient data delivery within the stipulated deadline, the sensor node which has sensed a delay-intolerant data needs to forward the data towards the mobile sink. In this paper, a Moore curve based trajectory of the mobile sink has been considered for efficient sink based data collection. This work is also motivated by anycast forwarding of delay sensitive data to the mobile sink before the sensed data lose its relevance, where the node follows a strict sleep-wake pattern. We have evaluated the effectiveness of the proposed trajectory based data collection scheme by simulation using MATLAB. The delay-



ScienceDirect

## Computers & Electrical Engineering

Volume 48, November 2015, Pages 417–435

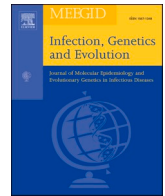
# An energy-efficient path determination strategy for mobile data collectors in wireless sensor network ☆

Nimisha Ghosh  , Indrajit Banerjee [Show more](#)  Share  Cite<https://doi.org/10.1016/j.compeleceng.2015.09.004>[Get rights and content](#)

## Abstract

In wireless sensor networks, introduction of mobility has been considered to be a good strategy to greatly reduce the energy dissipation of the static sensor nodes. This task is achieved by considering the path in which the mobile data collectors move to collect data from the sensors. In this work a data gathering approach is proposed in which some mobile collectors visit only certain sojourn points (SPs) or data collection points in place of all sensor nodes. The mobile collectors start out on their journey after gathering information about the network from the sink, gather data from the sensors and transfer the data to the sink. To address this problem, an algorithm named Mobile Collector Path Planning (MCP) is proposed. MCP schema is validated via computer simulation considering both obstacle free and obstacle-resisting network and based on metrics like energy consumption by the static sensor nodes and network life time. The simulation results show a reduction

Nimisha Ghosh



# An entropy-based study on mutational trajectory of SARS-CoV-2 in India

Daniele Santoni<sup>a,\*</sup>, Nimisha Ghosh<sup>b,c</sup>, Indrajit Saha<sup>d</sup>

<sup>a</sup> Institute for System Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, Via dei Taurini 19, Rome 00185, Italy

<sup>b</sup> Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland

<sup>c</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

<sup>d</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

## ARTICLE INFO

### Keywords:

COVID-19

Entropy

Hellinger distance

Mutation

SARS-CoV-2

## ABSTRACT

The pandemic of COVID-19 has been haunting us for almost the past two years. Although, the vaccination drive is in full swing throughout the world, different mutations of the SARS-CoV-2 virus are making it very difficult to put an end to the pandemic. The second wave in India, one of the worst sufferers of this pandemic, can be mainly attributed to the Delta variant i.e. B.1.617.2. Thus, it is very important to analyse and understand the mutational trajectory of SARS-CoV-2 through the study of the 26 virus proteins. In this regard, more than 17,000 protein sequences of Indian SARS-CoV-2 genomes are analysed using entropy-based approach in order to find the monthly mutational trajectory. Furthermore, Hellinger distance is also used to show the difference of the mutation events between the consecutive months for each of the 26 SARS-CoV-2 protein. The results show that the mutation rates and the mutation events of the viral proteins though changing in the initial months, start stabilizing later on for mainly the four structural proteins while the non-structural proteins mostly exhibit a more constant trend. As a consequence, it can be inferred that the evolution of the new mutative configurations will eventually reduce.

## 1. Introduction

Almost two years but COVID-19, the disease caused by SARS-CoV-2 is still disrupting our daily lives. Most of the cities around the globe including India have gone through various stages of lockdown to contain the spread of the virus. While the development and dissemination of vaccines have brought rays of hope, the circulation of the different variants of SARS-CoV-2 is still a cause of worry. As of now, the major variants of concern as declared by W.H.O are Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1) and Delta (B.1.617.2). Among these, the Delta variant was mainly responsible for the catastrophic second wave in India. Thus, mutations of SARS-CoV-2 need to be studied in order to understand its evolution. In this regard, Martin et al. (2021) studied Alpha, Beta and Gamma lineages to understand the evolutionary pattern of the virus. Dorp et al. (2020) curated 7666 SARS-CoV-2 genomes to analyse the genomic diversity of SARS-CoV-2, thereby identifying 198 filtered recurrent mutations. Yuan et al. (2020) performed global

analysis of 11,183 sequences to reveal the genetic diversity of SARS-CoV-2. Phylogenetic analysis was also carried out by Bai et al. (2020) with 16,373 SARS-CoV-2 genomes to reveal the evolution and molecular characteristics of SARS-CoV-2 while in Saha et al. (2021) the authors analysed more than 10,000 global SARS-CoV-2 genomes which identified 7209, 11,700, 119 and 53 unique mutation points as substitutions, deletions, insertions and SNPs. In Ghosh et al. (2021) performed phylogenetic analysis on more than 18,000 sequences to identify signature SNPs. Furthermore, in Saha et al. (2020a, 2020b), Saha et al. performed analysis of 566 Indian SARS-CoV-2 genomes to find the major mutation points in such genomes. On the other hand, entropy has been proven to be a very potent tool for the analysis of epidemics (Lucia et al., 2020). In Santos et al. (2021), Santos et al. proposed EntroPhylo which is an entropy based tool to select phylogenetic informative genetic regions and primer design. The tool considers the entropy value of each site and consequently the selected region is used for primer design. For evaluation purpose, EntroPhylo was used on the sequences of bovine

Nimisha Ghosh

\* Corresponding author.

E-mail address: [daniele.santoni@iasi.cnr.it](mailto:daniele.santoni@iasi.cnr.it) (D. Santoni).

<https://doi.org/10.1016/j.meegid.2021.105154>

Received 1 October 2021; Received in revised form 17 November 2021; Accepted 17 November 2021

Available online 19 November 2021

1567-1348/© 2021 The Authors.

Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

# Whole genome analysis of more than 10 000 SARS-CoV-2 virus unveils global genetic diversity and target region of NSP6

Indrajit Saha<sup>†</sup>, Nimisha Ghosh<sup>†</sup>, Ayan Pradhan, Nikhil Sharma, Debasree Maity and Kaushik Mitra

Corresponding author: Indrajit Saha, Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India. E-mail: [indrajit@nitttrkol.ac.in](mailto:indrajit@nitttrkol.ac.in)

<sup>†</sup>These authors contributed equally to this work

## Abstract

Whole genome analysis of SARS-CoV-2 is important to identify its genetic diversity. Moreover, accurate detection of SARS-CoV-2 is required for its correct diagnosis. To address these, first we have analysed publicly available 10 664 complete or near-complete SARS-CoV-2 genomes of 73 countries globally to find mutation points in the coding regions as substitution, deletion, insertion and single nucleotide polymorphism (SNP) globally and country wise. In this regard, multiple sequence alignment is performed in the presence of reference sequence from NCBI. Once the alignment is done, a consensus sequence is build to analyse each genomic sequence to identify the unique mutation points as substitutions, deletions, insertions and SNPs globally, thereby resulting in 7209, 11700, 119 and 53 such mutation points respectively. Second, in such categories, unique mutations for individual countries are determined with respect to other 72 countries. In case of India, unique 385, 867, 1 and 11 substitutions, deletions, insertions and SNPs are present in 566 SARS-CoV-2 genomes while 458, 1343, 8 and 52 mutation points in such categories are common with other countries. In majority (above 10%) of virus population, the most frequent and common mutation points between global excluding India and India are L37F, P323L, F506L, S507G, D614G and Q57H in NSP6, RdRp, Exon, Spike and ORF3a respectively. While for India, the other most frequent mutation points are T1198K, A97V, T315N and P13L in NSP3, RdRp, Spike and ORF8 respectively. These mutations are further visualised in protein structures and phylogenetic analysis has been done to show the diversity in virus genomes. Third, a web application is provided for searching mutation points globally and country wise. Finally, we have identified the potential conserved region as target that belongs to the coding region of ORF1ab, specifically to the NSP6 gene. Subsequently, we have provided the primers and probes using that conserved region so that it can be used for detecting SARS-CoV-2. **Contact:** [indrajit@nitttrkol.ac.in](mailto:indrajit@nitttrkol.ac.in) **Supplementary information:** Supplementary data are available at <http://www.nitttrkol.ac.in/indrajit/projects/COVID-Mutation-10K>

**Key words:** conserved region; multiple sequence alignment; NSP6; SNP; SARS-CoV-2; whole genome sequencing

**Indrajit Saha** Dr Saha is a faculty member in the Department of Computer Science and Engineering, NITTTR, Kolkata, India. His research interest includes computational intelligence, computational biology, machine learning, image processing and pattern recognition.

**Nimisha Ghosh** Dr Ghosh is a faculty member in the Department of Computer Science and Information Technology, ITER, Bhubaneswar, India. Her research interest includes computational intelligence, computational biology, machine learning, wireless sensor network and internet of things.

**Ayan Pradhan** Mr Pradhan is a faculty member in the Department of Computer Science and Engineering, Techno India University, Kolkata, India. His research interest includes optimization techniques, machine learning, artificial intelligence and data science.

**Nikhil Sharma** Mr Sharma is currently pursuing his B. Tech in the Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, India. His research interest includes computational biology and machine learning.

**Debasree Maity** Ms Maity is a faculty member in the Department of Electronics and Communication Engineering, MCKV Institute of Engineering, Howrah, India. Her research interest includes radiogenomics, population genetics and biosensor.

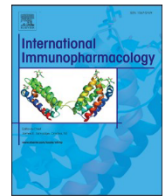
**Kaushik Mitra** Dr Mitra is presently working as an Associate Professor in the Department of Community Medicine, Burdwan Medical College, India. His area of expertise includes non-communicable diseases like diabetes, hypertension, cardiovascular diseases and cancers.

**Submitted:** 5 October 2020; **Received (in revised form):** 24 December 2020

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For Permissions, please email: [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

Nimisha Ghosh





# Bioinformatics pipeline unveils genetic variability to synthetic vaccine design for Indian SARS-CoV-2 genomes

Nimisha Ghosh<sup>a,1</sup>, Indrajit Saha<sup>b,\*</sup>, Nikhil Sharma<sup>c,1</sup>, Suman Nandi<sup>b</sup>

<sup>a</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

<sup>b</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

<sup>c</sup> Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

## ARTICLE INFO

### Keywords:

Bioinformatics Pipeline  
Clade  
Conserved Regions  
Non-synonymous signature SNP  
SARS-CoV-2  
T-cell epitopes

## ABSTRACT

In the worrisome scenarios of various waves of SARS-CoV-2 pandemic, a comprehensive bioinformatics pipeline is essential to analyse the virus genomes in order to understand its evolution, thereby identifying mutations as signature SNPs, conserved regions and subsequently to design epitope based synthetic vaccine. We have thus performed multiple sequence alignment of 4996 Indian SARS-CoV-2 genomes as a case study using MAFFT followed by phylogenetic analysis using Nextstrain to identify virus clades. Furthermore, based on the entropy of each genomic coordinate of the aligned sequences, conserved regions are identified. After refinement of the conserved regions, based on its length, one conserved region is identified for which the primers and probes are reported for virus detection. The refined conserved regions are also used to identify T-cell and B-cell epitopes along with their immunogenic and antigenic scores. Such scores are used for selecting the most immunogenic and antigenic epitopes. By executing this pipeline, 40 unique signature SNPs are identified resulting in 23 non-synonymous signature SNPs which provide 28 amino acid changes in protein. On the other hand, 12 conserved regions are selected based on refinement criteria out of which one is selected as the potential target for virus detection. Additionally, 22 MHC-I and 21 MHC-II restricted T-cell epitopes with 10 unique HLA alleles each and 17 B-cell epitopes are obtained for 12 conserved regions. All the results are validated both quantitatively and qualitatively which show that from genetic variability to synthetic vaccine design, the proposed pipeline can be used effectively to combat SARS-CoV-2.

## 1. Introduction

More than two years ago, SARS-CoV-2 put a massive halt to the freedom of human movement due to its high transmission rates [1]. Early study established the fact that SARS-CoV-2 virus is highly similar to that of the SARS-CoV-1 (95%–100%) [2]. In April 2021, India registered its second sudden surge in official cases with the Delta (B.1.617.2) variant. In late 2021, the third wave hit the country which was led by Omicron. Though, India is pushing towards a very large vaccination drive, concerns over the efficacy of the vaccine for such aggressive mutations are also increasing. Meanwhile, India is not the only country which has witnessed the new mutation strain of the evolving virus, variants in South Africa (501Y.V2) [3], United Kingdom (B.1.1.7) [4], Japan (E484K) [5], Brazil (P.1) [5] are also making their rounds. The

latest variant to join the bandwagon is Omicron (B.1.1.529). Although, previously it was suggested that such mutants are not going to affect the effectiveness of the vaccines currently in use, the emergence of Omicron has changed the equation. Moreover, new variants can affect the diagnosing procedure such as primer identification or antibody binding in RT-PCR. Ascoli [6] also suggested that a mutation in the Spike region of SARS-CoV-2 may affect the diagnosing procedure with greatest impact along with the increasing infection rates, transmissibility or even impacting people of younger age.

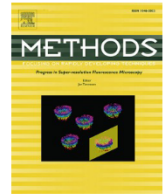
In the current scenario, it is an important and urgent task to study the frequently occurring mutations within the virus. In this regard, Yuan et al. [7] have analysed 11,183 SARS-CoV-2 genome from around the globe to identify the SNPs and critical SNPs with specific high mutation frequency along with the geographical pattern analysis. Further, they

\* Corresponding author.

E-mail address: [indrajit@nitttrkol.ac.in](mailto:indrajit@nitttrkol.ac.in) (I. Saha).

<sup>1</sup> Equally contributed

Nimisha Ghosh



# Characterisation of SARS-CoV-2 clades based on signature SNPs unveils continuous evolution

Nimisha Ghosh<sup>a,b,1</sup>, Indrajit Saha<sup>c,\*</sup>, Suman Nandi<sup>c,1</sup>, Nikhil Sharma<sup>d</sup>

<sup>a</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

<sup>b</sup> Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland

<sup>c</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

<sup>d</sup> Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

## ARTICLE INFO

### Keywords:

COVID-19

Clade

Deleterious mutations

Non-Synonymous Signature SNP

SARS-CoV-2

## ABSTRACT

Since the emergence of SARS-CoV-2 in Wuhan, China more than a year ago, it has spread across the world in a very short span of time. Although, different forms of vaccines are being rolled out for vaccination programs around the globe, the mutation of the virus is still a cause of concern among the research communities. Hence, it is important to study the constantly evolving virus and its strains in order to provide a much more stable form of cure. This fact motivated us to conduct this research where we have initially carried out multiple sequence alignment of 15359 and 3033 global dataset without Indian and the dataset of exclusive Indian SARS-CoV-2 genomes respectively, using MAFFT. Subsequently, phylogenetic analyses are performed using Nextstrain to identify virus clades. Consequently, the virus strains are found to be distributed among 5 major clades or clusters viz. 19A, 19B, 20A, 20B and 20C. Thereafter, mutation points as SNPs are identified in each clade. Henceforth, from each clade top 10 signature SNPs are identified based on their frequency i.e. number of occurrences in the virus genome. As a result, 50 such signature SNPs are individually identified for global dataset without Indian and dataset of exclusive Indian SARS-CoV-2 genomes respectively. Out of each 50 signature SNPs, 39 and 41 unique SNPs are identified among which 25 non-synonymous signature SNPs (out of 39) resulted in 30 amino acid changes in protein while 27 changes in amino acid are identified from 22 non-synonymous signature SNPs (out of 41). These 30 and 27 amino acid changes for the non-synonymous signature SNPs are visualised in their respective protein structure as well. Finally, in order to judge the characteristics of the identified clades, the non-synonymous signature SNPs are considered to evaluate the changes in proteins as biological functions with the sequences using PROVEAN and PolyPhen-2 while I-Mutant 2.0 is used to evaluate their structural stability. As a consequence, for global dataset without Indian sequences, G251V in ORF3a in clade 19A, F308Y and G196V in NSP4 and ORF3a in 19B are the unique amino acid changes which are responsible for defining each clade as they are all deleterious and unstable. Such changes which are common for both global dataset without Indian and dataset of exclusive Indian sequences are R203M in Nucleocapsid for 20B, T85I and Q57H in NSP2 and ORF3a respectively for 20C while for exclusive Indian sequences such unique changes are A97V in RdRp, G339S and G339C in NSP2 in 19A and Q57H in ORF3a in 20A.

## 1. Introduction

The first case of SARS-CoV-2 was registered in Wuhan China, 2019 and it quickly took over the normal functioning of human lives. In the initial phases, lock-down was implemented to limit the spread of infection. It is well known that virus mutations take place in the form of

single nucleotide variants, deletions and large structural variants [1] mainly due to replication and some hotspot mutations having severe impact on the host. Many cities around the globe have gone through staggered phases of lockdown in order to avoid the spread of the different strains of SARS-CoV-2. Among these strains, B.1.1.7 (Alpha) is found to be highly transmissible [2] and causes more severe pathogenic

\* Corresponding author.

E-mail address: [indrajit@nitttrkol.ac.in](mailto:indrajit@nitttrkol.ac.in) (I. Saha).

<sup>1</sup> Equally contributed.

Nimisha Ghosh





# Interactome of human and SARS-CoV-2 proteins to identify human hub proteins associated with comorbidities

Nimisha Ghosh<sup>a,b,1</sup>, Indrajit Saha<sup>c,\*</sup>, Nikhil Sharma<sup>d</sup>

<sup>a</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to Be University), Bhubaneswar, Odisha, India

<sup>b</sup> Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland

<sup>c</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

<sup>d</sup> Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

## ARTICLE INFO

### Keywords:

COVID-19

Comorbidities

Hub proteins

Protein-protein interaction

Repurposable drugs

SARS-CoV-2

## ABSTRACT

SARS-CoV-2 has a higher chance of progression in adults of any age with certain underlying health conditions or comorbidities like cancer, neurological diseases and in certain cases may even lead to death. Like other viruses, SARS-CoV-2 also interacts with host proteins to pave its entry into host cells. Therefore, to understand the behaviour of SARS-CoV-2 and design of effective antiviral drugs, host-virus protein-protein interactions (PPIs) can be very useful. In this regard, we have initially created a human-SARS-CoV-2 PPI database from existing works in the literature which has resulted in 7085 unique PPIs. Subsequently, we have identified at most 10 proteins with highest degrees viz. hub proteins from interacting human proteins for individual virus protein. The identification of these hub proteins is important as they are connected to most of the other human proteins. Consequently, when they get affected, the potential diseases are triggered in the corresponding pathways, thereby leading to comorbidities. Furthermore, the biological significance of the identified hub proteins is shown using KEGG pathway and GO enrichment analysis. KEGG pathway analysis is also essential for identifying the pathways leading to comorbidities. Among others, SARS-CoV-2 proteins viz. NSP2, NSP5, Envelope and ORF10 interacting with human hub proteins like COX4I1, COX5A, COX5B, NDUFS1, CANX, HSP90AA1 and TP53 lead to comorbidities. Such comorbidities are Alzheimer, Parkinson, Huntington, HTLV-1 infection, prostate cancer and viral carcinogenesis. Subsequently, using Enrichr tool possible repurposable drugs which target the human hub proteins are reported in this paper as well. Therefore, this work provides a consolidated study for human-SARS-CoV-2 protein interactions to understand the relationship between comorbidity and hub proteins so that it may pave the way for the development of anti-viral drugs.

## 1. Introduction

SARS-CoV-2, the virus responsible for COVID-19 has disrupted our daily lives and even after almost two years, we are still struggling in our fight against the virus. Though it originated in China, in a short time COVID-19 cases were reported from all around the globe. By September 2021, more than 229 million people have been affected by this virus with more than 4 million deaths.<sup>2</sup> The usual symptoms of COVID-19 range from common cough and cold, shortness of breath, fever to multiple organ failure which may eventually lead to death. Since this is a

RNA virus, it shows high mutations and new strains of the virus are also in circulation right now. According to W.H.O,<sup>3</sup> the strains of the virus declared as variants of concern are Alpha or B.1.1.7, Beta or B.1.351, Gamma or P.1 and Delta or B.1.617.2 [1–3].

SARS-CoV-2 encompasses four structural proteins, spike glycoprotein, envelope, membrane glycoprotein and nucleocapsid, apart from non-structural proteins (NSP1-NSP16) and accessory proteins like ORF3a, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF9c and ORF10 [4]. Viruses are incapable of living and reproducing outside a host body. Thus, they need to infiltrate a host for their survival. Protein-protein

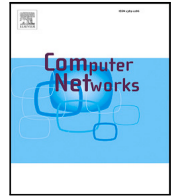
\* Corresponding author.

E-mail address: [indrajit@nittrkol.ac.in](mailto:indrajit@nittrkol.ac.in) (I. Saha).

<sup>1</sup> Equally contributed.

<sup>2</sup> <https://www.worldometers.info/coronavirus/>.

<sup>3</sup> <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>.



# Blockchain based secure smart city architecture using low resource IoTs

Rourab Paul <sup>a,\*</sup>, Nimisha Ghosh <sup>a</sup>, Suman Sau <sup>a</sup>, Amlan Chakrabarti <sup>b</sup>, Prasant Mohapatra <sup>c</sup>

<sup>a</sup> Siksha O Anusandhan, Bhubaneswar, India

<sup>b</sup> University of Calcutta, India

<sup>c</sup> University of California, Davis, United States of America

## ARTICLE INFO

### Keywords:

Access control  
Blockchain  
Cryptography  
IoT and smart city

## ABSTRACT

Internet of Things (IoT) is growing exponentially in research and industry. Although, many standard and conventional security solutions have been provided for IoT, it suffers from many privacy and security concerns. Standard security protocols are not suitable for majority of IoT devices because of its decentralized topology and resource-constraints. Blockchain (BC) finds its efficient application in IoT to preserve the five basic cryptographic primitives, such as confidentiality, authenticity, integrity, availability and non-repudiation. Adoption of conventional BC in IoT causes high energy consumption, delay and computational overheads which are not appropriate for various resource constrained IoT devices. To mitigate these problems, this work proposes a smart access control framework in a public and a private BC for a smart city application which makes it more efficient and secure as compared to the existing IoT applications. The proposed IoT based smart city architecture adopts BC technology for preserving all the cryptographic security and privacy issues. Moreover, proposed BC has minimal overhead as well. This work investigates the existing threat models and critical access control issues which handle multiple permissions of various processing nodes of IoT environment and detects relevant inconsistencies. Comparison in terms of all security issues with existing literature shows that the proposed architecture is competitively efficient in terms of security access control. The primary goal of this research article is to explore the possibility of BC as an alternative to standard security solutions for low resource IoT applications.

## 1. Introduction

The rapid growth of communication technology, network technology and increasing number of smart devices have made IoT a very relevant topic for research exploration. Moreover, existing literature survey shows that an IoT environment still suffers from privacy, security vulnerabilities, centralization and resource [1] issues despite some kind of existing security solutions. Conventional security solutions suffer from three significant issues [2–4] which are summarized as below:

### IoT Issues 1.1.

- (I) *The existing IoT applications use centralized communication (broker model) where all the nodes are authenticated, identified and connected through cloud server. A point failure in cloud server causes disruption of the entire network. These models are inappropriate when huge number of devices are connected [5].*

- (II) *The standard security solutions in IoT use third party for establishing trust. Several incidents are reported on unauthorized digital certificates issued by third parties [6].*
- (III) *In IoT, many devices like sensor nodes, smart devices and wearable devices are constrained in terms of computation, resources, power and bandwidth which makes it incompatible for existing complex security solutions [5].*

As mentioned in 1.1(I) and 1.1(II), the decentralized security, third party trust and privacy issues of IoT applications can be mitigated by conventional blockchain technology. However, conventional blockchain approaches cannot be adopted in IoT due to the following reasons:

- The consensus algorithms used in BC are computationally expensive which is unrealistic for IoT devices.
- The confirmation of transaction process of conventional BC requires significant delay whereas most of the IoT applications have strict deadlines.

\* Corresponding author.

E-mail addresses: [rourabpaul@soa.ac.in](mailto:rourabpaul@soa.ac.in) (R. Paul), [nimishaghosh@soa.ac.in](mailto:nimishaghosh@soa.ac.in) (N. Ghosh), [sumansau@soa.ac.in](mailto:sumansau@soa.ac.in) (S. Sau), [acakcs@caluniv.ac.in](mailto:acakcs@caluniv.ac.in) (A. Chakrabarti), [pmohapatra@ucdavis.edu](mailto:pmohapatra@ucdavis.edu) (P. Mohapatra).

<https://doi.org/10.1016/j.comnet.2021.108234>

Received 21 October 2020; Received in revised form 5 February 2021; Accepted 9 June 2021

Available online 19 June 2021

1389-1286/© 2021 Elsevier B.V. All rights reserved.

Nimisha Ghosh



# COVID-DeepPredictor: Recurrent Neural Network to Predict SARS-CoV-2 and Other Pathogenic Viruses

Indrajit Saha<sup>1\*†</sup>, Nimisha Ghosh<sup>2†</sup>, Debasree Maity<sup>3</sup>, Arijit Seal<sup>4</sup> and Dariusz Plewczynski<sup>5,6</sup>

<sup>1</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, India, <sup>2</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to Be University), Bhubaneswar, India, <sup>3</sup> Department of Electronics and Communication Engineering, MCKV Institute of Engineering, Howrah, India, <sup>4</sup> Cognizant Technology Solutions Pvt. Ltd., Kolkata, India, <sup>5</sup> Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland, <sup>6</sup> Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Warsaw, Poland

## OPEN ACCESS

### Edited by:

Xian-Tao Zeng,  
Wuhan University, China

### Reviewed by:

Sarath Chandra Janga,  
Indiana University, Purdue University  
Indianapolis, United States  
Xue-Qun Ren,  
Henan University, China

### \*Correspondence:

Indrajit Saha  
indrajit@nittrkol.ac.in

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

Received: 03 June 2020

Accepted: 13 January 2021

Published: 11 February 2021

### Citation:

Saha I, Ghosh N, Maity D, Seal A and  
Plewczynski D (2021)  
COVID-DeepPredictor: Recurrent  
Neural Network to Predict  
SARS-CoV-2 and Other Pathogenic  
Viruses. *Front. Genet.* 12:569120.  
doi: 10.3389/fgene.2021.569120

The COVID-19 disease for Novel coronavirus (SARS-CoV-2) has turned out to be a global pandemic. The high transmission rate of this pathogenic virus demands an early prediction and proper identification for the subsequent treatment. However, polymorphic nature of this virus allows it to adapt and sustain in different kinds of environment which makes it difficult to predict. On the other hand, there are other pathogens like SARS-CoV-1, MERS-CoV, Ebola, Dengue, and Influenza as well, so that a predictor is highly required to distinguish them with the use of their genomic information. To mitigate this problem, in this work COVID-DeepPredictor is proposed on the framework of deep learning to identify an unknown sequence of these pathogens. COVID-DeepPredictor uses Long Short Term Memory as Recurrent Neural Network for the underlying prediction with an alignment-free technique. In this regard,  $k$ -mer technique is applied to create Bag-of-Descriptors (BoDs) in order to generate Bag-of-Unique-Descriptors (BoUDs) as vocabulary and subsequently embedded representation is prepared for the given virus sequences. This predictor is not only validated for the dataset using  $K$ -fold cross-validation but also for unseen test datasets of SARS-CoV-2 sequences and sequences from other viruses as well. To verify the efficacy of COVID-DeepPredictor, it has been compared with other state-of-the-art prediction techniques based on Linear Discriminant Analysis, Random Forests, and Gradient Boosting Method. COVID-DeepPredictor achieves 100% prediction accuracy on validation dataset while on test datasets, the accuracy ranges from 99.51 to 99.94%. It shows superior results over other prediction techniques as well. In addition to this, accuracy and runtime of COVID-DeepPredictor are considered simultaneously to determine the value of  $k$  in  $k$ -mer, a comparative study among  $k$  values in  $k$ -mer, Bag-of-Descriptors (BoDs), and Bag-of-Unique-Descriptors (BoUDs) and a comparison between COVID-DeepPredictor and Nucleotide BLAST have also been performed. The code, training, and test datasets used for COVID-DeepPredictor are available at <http://www.nittrkol.ac.in/indrajit/projects/COVID-DeepPredictor/>.

**Keywords:** long-short term memory, SARS-CoV-2, sequence analysis, virus prediction, genomic information

Nimisha Ghosh

[Published: 19 August 2019](#)

# Differential Evolution and Mobile Sink Based On-Demand Clustering Protocol for Wireless Sensor Network

[Nimisha Ghosh](#) , [Tripti Prasad](#) & [Indrajit Banerjee](#)

[Wireless Personal Communications](#) **109**, 1875–1895 (2019)

**144** Accesses | **3** Citations | [Metrics](#)

## Abstract

Efficient cluster formation is an important aspect of wireless sensor networks. But to transfer the data from the cluster heads to the static sink may lead to an energy hole problem where the cluster heads near the sink deplete their energy faster than those away from the sink. Using a mobile sink helps in

Nimisha Ghosh



# Fault Matters: Sensor data fusion for detection of faults using Dempster–Shafer theory of evidence in IoT-based applications

Nimisha Ghosh <sup>a,\*</sup>, Rourab Paul <sup>b</sup>, Satyabrata Maity <sup>b</sup>, Krishanu Maity <sup>a</sup>, Sayantan Saha <sup>b</sup>

<sup>a</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

<sup>b</sup> Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

## ARTICLE INFO

### Keywords:

Classification  
Data fusion  
Dempster–Shafer  
Fault detection  
Internet of Things (IoT)

## ABSTRACT

Fault detection in sensor nodes is a pertinent issue that has been an important area of research for a very long time. But it is not explored much as yet in the context of Internet of Things. Internet of Things work with a massive amount of data so the responsibility for guaranteeing the accuracy of the data also lies with it. Moreover, a lot of important and critical decisions are made based on these data, so ensuring its correctness and accuracy is also very important. Also, the detection needs to be as precise as possible to avoid negative alerts. For this purpose, this work has adopted Dempster–Shafer theory of evidence to collate the information from sensors to come up with a decision regarding the faulty status of a sensor node from a data-centric perspective. To verify the validity of the proposed method, simulations have been performed on a benchmark data set and data collected through a test bed in a laboratory set-up. For the different types of faults, the proposed method shows very high accuracy for both the benchmark (99.8%) and laboratory data sets (99.9%) when compared to the other state-of-the-art machine learning techniques.

## 1. Introduction

Internet of Things (IoT) have extended computing capability to everyday objects which are embedded with sensors to come up with a huge repository of data which can be exchanged and consumed without any human intervention. This data plays a huge role in decision-making policies. Thus, the foremost requirement in any application of IoT would be to ensure the reliability of data. For this purpose, detection of a faulty device which produces anomalous data is very important. A data is considered to be faulty (anomalous or outlier) if it deviates significantly from the normal range of value (Hawkins, 1980). These deviations may turn out to be catastrophic if not detected on time. Faulty sensor data is attributed to the crash of Lion Air 737 on 29th October 2018 which killed all 189 people on-board. Erroneous signals from sensor caused continuous nose-down motions resulting in the ill-fated crash.

Outlier detection is a very pertinent topic in any field which is concerned with data like fraud and intrusion detection (Hajiheidari et al., 2019), weather monitoring (Shepherd & Burian, 2003), traffic anomaly detection (Xie et al., 2018), sensor faults in heating, ventilation and air-conditioning (HVAC) systems (Reppa et al., 2015) etc. In

an IoT application, monitoring and collection of real-time data from various sensors are necessary for the proper functioning of the system. All the data that is collected should be meticulously processed for the smooth operation of any system. The inherent components of any sensor network are prone to failures and this may result in incorrect readings. Like any other practical system, faulty sensor data may cause instability in IoT-based system as well. Thus, a proper fault detection technique needs to be incorporated to report the abnormality in the sensor data and thus identify a compromised node. Fault detection inherently can be considered to be a classification problem. Classification is a supervised learning technique in which labels are assigned to an observation based on its feature values. In a classification problem, a suitable learning technique is first trained with the data so that the classifier learns the difference between the different classes. Later, the classifier can be used to test a new observation to make a conclusion about its class. This work uses Dempster–Shafer Theory of Evidence (DSTE) (Dempster, 2008; Shafer, 1976) which is a mathematical model that combines information from different sources. DSTE deals with uncertainty and thus can work with missing informations and conflicting informations. Thus, it can be considered to be a more general approach

\* Corresponding author.

E-mail addresses: [nimishaghosh@soa.ac.in](mailto:nimishaghosh@soa.ac.in) (N. Ghosh), [rourabpaul@soa.ac.in](mailto:rourabpaul@soa.ac.in) (R. Paul), [satyabratamaity@soa.ac.in](mailto:satyabratamaity@soa.ac.in) (S. Maity), [krishanumaity@soa.ac.in](mailto:krishanumaity@soa.ac.in) (K. Maity), [sayantansaha@soa.ac.in](mailto:sayantansaha@soa.ac.in) (S. Saha).

<https://doi.org/10.1016/j.eswa.2020.113887>

Received 8 April 2020; Received in revised form 28 July 2020; Accepted 13 August 2020

Available online 22 August 2020

0957-4174/© 2020 Elsevier Ltd. All rights reserved.

Nimisha Ghosh





## Research Paper

## Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP



Indrajit Saha<sup>a,\*</sup>, Nimisha Ghosh<sup>b,1</sup>, Debasree Maity<sup>c</sup>, Nikhil Sharma<sup>d</sup>,  
Jnanendra Prasad Sarkar<sup>e,f</sup>, Kaushik Mitra<sup>g</sup>

<sup>a</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

<sup>b</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Orissa, India

<sup>c</sup> Department of Electronics and Communication Engineering, MCKV Institute of Engineering, Howrah, West Bengal, India

<sup>d</sup> Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

<sup>e</sup> Larsen & Toubro Infotech, Pune, India

<sup>f</sup> Department of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal, India

<sup>g</sup> Department of Community Medicine, Burdwan Medical College, Bardhaman, West Bengal, India

## ARTICLE INFO

## Keywords:

Multiple sequence alignment

Point mutation

SNPs

SARS-CoV-2

## ABSTRACT

The wave of COVID-19 is a big threat to the human population. Presently, the world is going through different phases of lock down in order to stop this wave of pandemic; India being no exception. We have also started the lock down on 23rd March 2020. In this current situation, apart from social distancing only a vaccine can be the proper solution to serve the population of human being. Thus it is important for all the nations to perform the genome-wide analysis in order to identify the genetic variation in Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) so that proper vaccine can be designed. This fast motivated us to analyze publicly available 566 Indian complete or near complete SARS-CoV-2 genomes to find the mutation points as substitution, deletion and insertion. In this regard, we have performed the multiple sequence alignment in presence of reference sequence from NCBI. After the alignment, a consensus sequence is built to analyze each genome in order to identify the mutation points. As a consequence, we have found 933 substitutions, 2449 deletions and 2 insertions, in total 3384 unique mutation points, in 566 genomes across 29.9 K bp. Further, it has been classified into three groups as 100 clusters of mutations (mostly deletions), 1609 point mutations as substitution, deletion and insertion and 64 SNPs. These outcomes are visualized using BioCircos and bar plots as well as plotting entropy value of each genomic location. Moreover, phylogenetic analysis has also been performed to see the evolution of SARS-CoV-2 virus in India. It also shows the wide variation in tree which indeed vivid in genomic analysis. Finally, these SNPs can be the useful target for virus classification, designing and defining the effective dose of vaccine for the heterogeneous population.

## 1. Introduction

Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) and its associated disease known as COVID-19, was originated in Wuhan, China (Zhu et al., 2020). It has wreaked havoc on human lives and declared as a pandemic by World Health Organisation on 11th March 2020. Among others, symptoms of COVID-19 include fever, cough and shortness of breath (Chen et al., 2020). In more severe cases, infection may lead to pneumonia (Zhou et al., 2020), kidney failure and eventual death. As of now, no vaccine or medicine has been invented or

discovered and the only protective measures are being taken by different countries are through lock downs and social distancing. However, even these extreme measures have not been able to contain the spread of SARS-CoV-2. Everyday thousands of new cases are coming into light. According to the record of 27th June, globally more than 9.8 million people are affected by this deadly virus, with a total death count close to 500 thousand (Worldometer, 2020). India is also fighting hard to keep the citizens safe through lock down and other protective measures. Yet the virus has shown a surge in the country with more than 500 thousand affected cases and death cases of more than 15 thousand

\* Corresponding author.

E-mail address: [indrajit@nitttrkol.ac.in](mailto:indrajit@nitttrkol.ac.in) (I. Saha).

<sup>1</sup> Equally contributed.

<https://doi.org/10.1016/j.meegid.2020.104457>

Received 10 May 2020; Received in revised form 3 July 2020; Accepted 5 July 2020

Available online 11 July 2020

1567-1348/ © 2020 Elsevier B.V. All rights reserved.

Nimisha Ghosh



# Genome-wide analysis of Indian SARS-CoV-2 genomes to identify T-cell and B-cell epitopes from conserved regions based on immunogenicity and antigenicity

Nimisha Ghosh <sup>a,1</sup>, Nikhil Sharma <sup>b,1</sup>, Indrajit Saha <sup>c,\*</sup>, Sudipto Saha <sup>d</sup>

<sup>a</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Orissa, India

<sup>b</sup> Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

<sup>c</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

<sup>d</sup> Division of Bioinformatics Bose Institute, Kolkata, West Bengal, India

## ARTICLE INFO

### Keywords:

B-cell epitopes  
Conserved regions  
SARS-CoV-2  
Peptide based vaccine  
Physico-chemical properties  
T-cell epitopes

## ABSTRACT

SARS-CoV-2 has a high transmission rate and shows frequent mutations, thus making vaccine development an arduous task. However, researchers around the globe are working hard to find a solution e.g. synthetic vaccine. Here, we have performed genome-wide analysis of 566 Indian SARS-CoV-2 genomes to extract the potential conserved regions for identifying peptide based synthetic vaccines, viz. epitopes with high immunogenicity and antigenicity. In this regard, different multiple sequence alignment techniques are used to align the SARS-CoV-2 genomes separately. Subsequently, consensus conserved regions are identified after finding the conserved regions from each aligned result of alignment techniques. Further, the consensus conserved regions are refined considering that their lengths are greater than or equal to 60nt and their corresponding proteins are devoid of any stop codons. Subsequently, their specificity as query coverage are verified using Nucleotide BLAST. Finally, with these consensus conserved regions, T-cell and B-cell epitopes are identified based on their immunogenic and antigenic scores which are then used to rank the conserved regions. As a result, we have ranked 23 consensus conserved regions that are associated with different proteins. This ranking also resulted in 34 MHC-I and 37 MHC-II restricted T-cell epitopes with 16 and 19 unique HLA alleles and 29 B-cell epitopes. After ranking, the consensus conserved region from NSP3 gene is obtained that is highly immunogenic and antigenic. In order to judge the relevance of the identified epitopes, the physico-chemical properties and binding conformation of the MHC-I and MHC-II restricted T-cell epitopes are shown with respect to HLA alleles.

## 1. Introduction

In December 2019, China reported a sudden outbreak of pneumonia due to an unknown source in Hubei province, Wuhan city [1] which later got attributed to a virus named SARS-CoV-2. SARS-CoV-2 belongs to the family of Coronaviridae which also houses SARS-CoV-1 [2,3] and MERS-CoV [4] virus. Genomic sequence analysis of the newly reported virus was found to be highly similar to that of SARS-CoV (95%–100%), thus showing the evolutionary similarity between SARS-CoV and SARS-CoV-2 [5]. By October 2020, India has registered over 7.65 million cases [6], making it one of the most affected countries in the world. Symptoms of the COVID-19 vary from fever, cough, myalgia, dyspnoea and

diarrhoea to severe respiratory distress which may require life support systems. In severe cases, it may even lead to death [7]. Considering these consequences, World Health Organisation (WHO) suggested to interrupt human–human contact in the form of total lock downs along with precautionary measures such as face masks and hand sanitizers to control the spread of COVID-19. Hence, it is the need of the hour to find a cure for COVID-19 in the form of vaccine.

Classical methods of vaccine design like attenuation of the virus through external sources such as micro-organisms to mitigate its harm or virulence usually depends on the response of the virus itself. Sometimes mutations in the virus genome can result in autoimmune response eventually making the virus even more virulent. Hence, such classic

\* Corresponding author.

E-mail address: [indrajit@nittrkol.ac.in](mailto:indrajit@nittrkol.ac.in) (I. Saha).

<sup>1</sup> Equally contributed.

Received January 24, 2022, accepted February 16, 2022, date of publication February 25, 2022, date of current version March 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3154053

# Genome-Wide Analysis to Identify Palindromes, Mirror and Inverted Repeats in SARS-CoV-2, MERS-CoV and SARS-CoV-1

NIMISHA GHOSH<sup>1</sup>, INDRAJIT SAHA<sup>2</sup>, AND DARIUSZ PLEWCZYNSKI<sup>3,4</sup>

<sup>1</sup>Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha 751030, India

<sup>2</sup>Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal 700106, India

<sup>3</sup>Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-661 Warsaw, Poland

<sup>4</sup>Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, 02-097 Warsaw, Poland

Corresponding authors: Indrajit Saha (indrajit@nittrkol.ac.in) and Dariusz Plewczynski (d.plewczynski@mini.pw.edu.pl)

This work was supported in part by the Core Research Grant (CRG) Short Term Research Grant on COVID-19 (CVD/2020/000991) from the Science and Engineering Research Board (SERB), Department of Science and Technology, Government of India; in part by the Warsaw University of Technology within the Excellence Initiative: Research University (IDUB) Programme; and in part by the Polish National Science Centre under Grant 2019/35/O/ST6/02484 and Grant 2020/37/B/NZ2/03757. Computations were performed thanks to the Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology using Artificial Intelligence HPC platform financed by Polish Ministry of Science and Higher Education (decision no. 7054/IA/SP/2020 of 2020-08-28). The work was co-supported by EU-funded the Marie Skłodowska-Curie action (MSCA) Innovative Training Network (ITN) named Enhpathy (www.enhpathy.eu) 'Molecular Basis of Human enhanceropathies'.

**ABSTRACT** Research pertaining to SARS-CoV-2 is in full swing to understand the origin and evolution of this deadly virus that can lead to its rapid detection. To achieve this, atypical genomic sequences which may be unique to SARS-CoV-2 or Coronaviridae family in general may be investigated. Such sequences in virus genomes may be responsible for target prediction, replication, defence mechanisms and viral packaging. This fact has motivated us to explore the different types of repeats such as palindromes, mirror repeats and inverted repeats in SARS-CoV-2, MERS-CoV and SARS-CoV-1. For this purpose, the respective reference sequence of SARS-CoV-2, MERS-CoV and SARS-CoV-1 is divided into descriptors of sequences of length  $k$  using  $k$ -mer technique. Thereafter, these descriptors are represented as a collection of tokens which are subsequently used for the identification of palindrome, mirror repeat and inverted repeat in the respective reference sequence. The highest number of palindromes, mirror repeats and inverted repeats are identified for descriptor length 10. As a result, for palindromes such values are 38, 42 and 33 and for mirror repeats they are 52, 38 and 33 for SARS-CoV-2, MERS-CoV and SARS-CoV-1 respectively. For inverted repeats, with a descriptor length 10 and intervening length 5, the values are 59, 56 and 70 respectively. Moreover, the identified repeats are then searched for in 108246, 291 and 340 SARS-CoV-2, MERS-CoV and SARS-CoV-1 virus sequences respectively to find the population coverage of such repeats. It surpasses 99% in most cases and even 100% for some. Furthermore, GC contents which mostly lie between 20%-50% are evaluated for these repeats as well in order to understand their binding efficacy.

**INDEX TERMS** COVID-19, inverted repeats, mirror repeats, palindromes, SARS-CoV-2.

## I. INTRODUCTION

The outbreak of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2), the virus causing COVID-19, has caused more than five million deaths in different parts of the world leading to researches around the globe to control this virus. The most common symptoms of COVID-19 include fever, dry cough, dyspnea, headache and pneumonia and the less common symptoms are gastrointestinal

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott<sup>1</sup>.

symptoms, myalgias, rhinorrhea and chest pain [1]; the fatality rate is high in patients suffering from cancer, diabetes or cardiovascular disorders [2]. SARS-CoV-2 belongs to the family of Coronaviridae which also houses MERS-CoV and SARS-CoV-1 viruses [1], [3]. Research pertaining to SARS-CoV-2 is in full swing to understand the origin and evolution of this deadly virus. This may lead to the immune-related studies of viruses [4]. To achieve this, atypical genomic sequences which may be unique to SARS-CoV-2 or Coronaviridae family in general may be investigated [5]–[8]. Studies [9]–[11] have suggested that such sequences

Nimisha Ghosh



# Hotspot Mutations in SARS-CoV-2

Indrajit Saha<sup>1\*</sup>, Nimisha Ghosh<sup>2†</sup>, Nikhil Sharma<sup>3</sup> and Suman Nandi<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, India, <sup>2</sup>Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India, <sup>3</sup>Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, India

Since its emergence in Wuhan, China, severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) has spread very rapidly around the world, resulting in a global pandemic. Though the vaccination process has started, the number of COVID-affected patients is still quite large. Hence, an analysis of hotspot mutations of the different evolving virus strains needs to be carried out. In this regard, multiple sequence alignment of 71,038 SARS-CoV-2 genomes of 98 countries over the period from January 2020 to June 2021 is performed using MAFFT followed by phylogenetic analysis in order to visualize the virus evolution. These steps resulted in the identification of hotspot mutations as deletions and substitutions in the coding regions based on entropy greater than or equal to 0.3, leading to a total of 45 unique hotspot mutations. Moreover, 10,286 Indian sequences are considered from 71,038 global SARS-CoV-2 sequences as a demonstrative example that gives 52 unique hotspot mutations. Furthermore, the evolution of the hotspot mutations along with the mutations in variants of concern is visualized, and their characteristics are discussed as well. Also, for all the non-synonymous substitutions (missense mutations), the functional consequences of amino acid changes in the respective protein structures are calculated using PolyPhen-2 and I-Mutant 2.0. In addition to this, SSIPE is used to report the binding affinity between the receptor-binding domain of Spike protein and human ACE2 protein by considering L452R, T478K, E484Q, and N501Y hotspot mutations in that region.

**Keywords:** COVID-19, deletions, entropy, hotspot mutations, SARS-CoV-2 genomes, substitution

## OPEN ACCESS

### Edited by:

Yang Zhang,  
University of Michigan, United States

### Reviewed by:

Xiaoqiang Huang,  
University of Michigan, United States  
Yavuz Oktay,  
Dokuz Eylul University, Turkey

### \*Correspondence:

Indrajit Saha  
indrajit@nittrkol.ac.in

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Computational Genomics,  
a section of the journal  
Frontiers in Genetics

**Received:** 04 August 2021

**Accepted:** 07 October 2021

**Published:** 29 November 2021

### Citation:

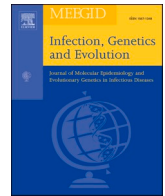
Saha I, Ghosh N, Sharma N and  
Nandi S (2021) Hotspot Mutations  
in SARS-CoV-2.  
Front. Genet. 12:753440.  
doi: 10.3389/fgene.2021.753440

## 1 INTRODUCTION

COVID-19 caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) was first identified in late December 2019 and has a high transmission rate (Zhu et al., 2020). The WHO declared this outbreak as a pandemic on March 11, 2020 (Cucinotta and Vanelli, 2020). Like other coronaviruses, SARS-CoV-2 is also an enveloped single-stranded RNA virus containing nearly 30 K nucleotide sequences (Alexandersen et al., 2020). SARS-CoV-2 encompasses 11 coding regions, which include ORF1ab, Spike (S), ORF3a, Envelope (E), Membrane (M), ORF6, ORF7a, ORF7b, ORF8, Nucleocapsid (N), and ORF10.

Though the vaccination process has started, the virus is evolving and spreading all across the world, causing fresh waves every few months. Since the virus is mutating frequently, it creates new variant of the original virus. Among several variants, B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma), and B.1.617.2 (Delta) are declared as variants of concern (Singh et al., 2021). In this regard, the variant B.1.1.7 was first identified in the United Kingdom, which contains E484K, N501Y, D614G, and P681H mutations in Spike glycoprotein (Tang et al., 2020). In December 2020, the variant

Nimisha Ghosh



## Research paper

# Immunogenicity and antigenicity based T-cell and B-cell epitopes identification from conserved regions of 10664 SARS-CoV-2 genomes

Nimisha Ghosh<sup>a,1</sup>, Nikhil Sharma<sup>b,1</sup>, Indrajit Saha<sup>c,\*</sup>

<sup>a</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

<sup>b</sup> Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

<sup>c</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

## ARTICLE INFO

## Keywords:

B-cell epitopes  
Conserved regions  
Epitopes  
SARS-CoV-2  
Synthetic vaccine  
T-cell epitopes

## ABSTRACT

The surge of SARS-CoV-2 has created a wave of pandemic around the globe due to its high transmission rate. To contain this virus, researchers are working around the clock for a solution in the form of vaccine. Due to the impact of this pandemic, the economy and healthcare have immensely suffered around the globe. Thus, an efficient vaccine design is the need of the hour. Moreover, to have a generalised vaccine for heterogeneous human population, the virus genomes from different countries should be considered. Thus, in this work, we have performed genome-wide analysis of 10,664 SARS-CoV-2 genomes of 73 countries around the globe in order to identify the potential conserved regions for the development of peptide based synthetic vaccine viz. epitopes with high immunogenic and antigenic scores. In this regard, multiple sequence alignment technique viz. Clustal Omega is used to align the 10,664 SARS-CoV-2 virus genomes. Thereafter, entropy is computed for each genomic coordinate of the aligned genomes. The entropy values are then used to find the conserved regions. These conserved regions are refined based on the criteria that their lengths should be greater than or equal to 60 nt and their corresponding protein sequences are without any stop codons. Furthermore, Nucleotide BLAST is used to verify the specificity of the conserved regions. As a result, we have obtained 17 conserved regions that belong to NSP3, NSP4, NSP6, NSP8, RdRp, Helicase, endoRNase, 2'-O-RMT, Spike glycoprotein, ORF3a protein, Membrane glycoprotein and Nucleocapsid protein. Finally, these conserved regions are used to identify the T-cell and B-cell epitopes with their corresponding immunogenic and antigenic scores. Based on these scores, the most immunogenic and antigenic epitopes are then selected for each of these 17 conserved regions. Hence, we have obtained 30 MHC-I and 24 MHC-II restricted T-cell epitopes with 14 and 13 unique HLA alleles and 21 B-cell epitopes for the 17 conserved regions. Moreover, for validating the relevance of these epitopes, the binding conformation of the MHC-I and MHC-II restricted T-cell epitopes are shown with respect to HLA alleles. Also, the physico-chemical properties of the epitopes are reported along with Ramchandran plots and Z-Scores and the population coverage is shown as well. Overall, the analysis shows that the identified epitopes can be considered as potential candidates for vaccine design.

## 1. Introduction

In late December 2019, China witnessed the rise of novel coronavirus, SARS-CoV-2, whose exponential rise took everyone by surprise due to its high transmission rate. Later, the origin of the virus was linked to coronaviridae family which also includes SARS-CoV-1 and MERS-CoV (Zhou et al., 2020). By the start of March, number of people from different countries around the globe were found to be infected with the

virus. Hence, W.H.O. declared it as a pandemic thereby forcing authorities to adopt counter measures to limit the spread of SARS-CoV-2 among the masses. People infected with SARS-CoV-2 can exhibit mild and moderate symptoms, ranging from asymptomatic to high body temperature along with cough, sore throat and may also pose life threatening situations in people suffering from other diseases such as diabetes, cardiovascular diseases etc.. Some of the most widely accepted methods to curb the spread of the virus were nationwide lock down,

\* Corresponding author.

E-mail address: [indrajit@nittrkol.ac.in](mailto:indrajit@nittrkol.ac.in) (I. Saha).

<sup>1</sup> Equally contributed.

Nimisha Ghosh





## iDCR: Improved Dempster Combination Rule for multisensor fault diagnosis

Nimisha Ghosh<sup>a,\*</sup>, Sayantan Saha<sup>b</sup>, Rourab Paul<sup>b</sup><sup>a</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India<sup>b</sup> Department of Computer Science and Engineering, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

## ARTICLE INFO

## Keywords:

Conflict evidence  
 Dempster-Shafer theory of evidence (DSTE)  
 Dempster's Combination Rule (DCR)  
 Fault diagnosis  
 Weighted Deng entropy

## ABSTRACT

Data gathered from multiple sensors can be effectively fused for accurate monitoring of many engineering applications. In the last few years, one of the most sought after applications for multisensor fusion has been fault diagnosis. Dempster-Shafer Theory of Evidence along with Dempster's Combination Rule is a very popular method for multisensor fusion which can be successfully applied to fault diagnosis. But if the information obtained from the different sensors shows high conflict, the classical Dempster's Combination Rule may produce counter-intuitive result. To overcome this shortcoming, this paper proposes an improved combination rule for multisensor data fusion. Numerical examples have been put forward to show the effectiveness of the proposed method. Comparative analysis has also been carried out with existing methods to show the superiority of the proposed method in multisensor fault diagnosis.

## 1. Introduction

With the growing demands of surveillance systems and the booming market of Internet of Things, sensors have become an imperative part of everyday life. Thus, the reliability on sensors is increasing with each passing day. However, traditional single sensor systems have not really been able to match up to the sensor performance. For instance, if with three sensors it is attempted to detect the fault in a machine, where the first one shows a temperature of 35°C, second one 42°C and the third one 37°C, the machine is considered as faulty if the temperature is above 40°C. So, it is difficult to judge from the sensor values if the machine can be considered as faulty. If information from only one sensor is taken into account, then there is high probability of information loss and further varied decision results will be acquired. This can be attributed to the conflicting results of the sensors. In reality, the information from different sensors may be inaccurate, fuzzy and conflicting, resulting in a difficult decision making. To mitigate these problem, Bayesian theory (Cai et al., 2014), fuzzy sets (Mofizul Islam et al., 2000), Z-numbers (Jiang et al., 2016b), D-numbers (Liu et al., 2014), evidence theory etc. have been put forth. In Cai et al. (2014), to build a fault diagnosis model, the authors have used a Bayesian network with two layers namely a fault layer and a fault symptom layer. In Mofizul Islam et al. (2000), fuzzy set theory has been applied for fault diagnosis. To combat data uncertainty, Jiang et al. (2016b) has used Z-number to combine the data from different sensors. The authors have used D-number theory in Liu et al. (2014) to handle uncertain information for risk analysis. This work has shown that D-number is

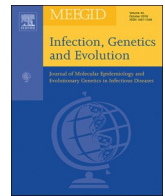
effective in combining data from different sensors. But for the error in the sensor data, the information may result in high conflict, thus culminating in error prone fusion result. Dempster-Shafer evidence theory can deal with this conflicting information.

Dempster-Shafer theory of evidence (DSTE) (Dempster, 2008; Shafer, 1976) is a mathematical concept for combining data from different sensors. It is a generalisation of Bayesian method and offers better fusion result than most of the aforementioned methods. Thus, evidence theory has been incorporated in many applications to deal with incomplete and uncertain information. An agent-oriented intelligent fault diagnosis system has been proposed in Luo et al. (2012). Belief entropy has been used in Yuan et al. (2016) by considering sensor reliability and assigning different weights using a distance function. The authors in Jiang et al. (2016a) have used Deng's entropy to measure and modify the evidence by using discounting coefficients. Finally, Dempster's Combination Rule (DCR) has been used to fuse the modified evidence to detect fault in a motor rotor. A novel fault diagnosis method of proton exchange membrane fuel cell (PEMFC) system has been proposed in Liu et al. (2019) which combines Extreme Learning Machine (ELM) and Dempster-Shafer theory of evidence. The failure diagnosis model is taken care of by ELM and the diagnostic output is combined by DSTE. Support vector machine (SVM)-evidence theory model has been used in Hui et al. (2017) to solve information gathered by every SVM model to increase the accuracy of classification. The resulting model has been shown to improve the accuracy by removing all conflicting information from the original SVM model. Thus, it can

\* Corresponding author.

E-mail addresses: [nimishaghosh@soa.ac.in](mailto:nimishaghosh@soa.ac.in) (N. Ghosh), [sayantansaha@soa.ac.in](mailto:sayantansaha@soa.ac.in) (S. Saha), [rourabpaul@soa.ac.in](mailto:rourabpaul@soa.ac.in) (R. Paul).

Nimisha Ghosh



## Short communication

## Inferring the genetic variability in Indian SARS-CoV-2 genomes using consensus of multiple sequence alignment techniques

Indrajit Saha<sup>a,\*</sup>, Nimisha Ghosh<sup>b,1</sup>, Debasree Maity<sup>c</sup>, Nikhil Sharma<sup>d</sup>, Kaushik Mitra<sup>e</sup><sup>a</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India<sup>b</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India<sup>c</sup> Department of Electronics and Communication Engineering, MCKV Institute of Engineering, Howrah, West Bengal, India<sup>d</sup> Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India<sup>e</sup> Department of Community Medicine, Burdwan Medical College, Burdwan, West Bengal, India

## ARTICLE INFO

## Keywords:

Multiple sequence alignment

Point mutation

SNP

SARS-CoV-2

## ABSTRACT

Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is a threat to the human population and has created a worldwide pandemic. Daily thousands of people are getting affected by the SARS-CoV-2 virus; India being no exception. In this situation, there is no doubt that vaccine is the primary prevention strategy to contain the wave of COVID-19 pandemic. In this regard, genome-wide analysis of SARS-CoV-2 is important to understand its genetic variability. This has motivated us to analyse 566 Indian SARS-CoV-2 sequences using multiple sequence alignment techniques viz. ClustalW, MUSCLE, ClustalO and MAFFT to align and subsequently identify the lists of mutations as substitution, deletion, insertion and SNP. Thereafter, a consensus of these results, called as Consensus Multiple Sequence Alignment (CMSA), is prepared to have the final list of mutations so that the advantages of all four alignment techniques can be preserved. The analysis shows 767, 2025 and 54 unique substitutions, deletions and SNPs in Indian SARS-CoV-2 genomes. More precisely, out of 54 SNPs, 4 SNPs are present close to the 60% of the virus population. The results of this experiment can be useful for virus classification, designing and defining the dose of vaccine for the Indian population.

## 1. Introduction

Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) causes a disease COVID-19 which originated in Wuhan, China (Zhu et al., 2020). This disease has brought a wave of pandemic throughout the world. In this worrisome situation, undoubtedly vaccine is the primary prevention strategy to contain this virus. However, the development process of vaccine is time consuming and requires the analysis of genetic variability of the virus population so that effective and safe vaccine can be developed for heterogeneous human population (Poland, 2020). In this regard, Tung Phan has performed a genomic analysis to show the evolution of SARS-CoV-2 (Phan, 2020), subsequently, we have also analysed Indian SARS-CoV-2 genomes in (Saha et al., 2020). In continuation to these works, we have further conducted the genomic analysis of 566 SARS-CoV-2 virus population to identify the mutation as substitution, deletion, insertion and Single Nucleotide Polymorphism (SNP). Generally, 1% of the population which is being affected by substitution is termed as SNP. We have concerned ourselves

with only non-synonymous mutations as they are responsible for changes in the amino acids. In order to find the genetic variability, the multiple sequence alignment (MSA) is very much essential in presence of reference sequence. On the other hand, it is also known that the multiple sequence alignment technique provides a near optimal result. Therefore, different alignment techniques can provide different results on the same pool of sequences. Hence, we have used four different well-known multiple sequence alignment techniques viz. ClustalW (Thompson et al., 1994), MUSCLE (Edgar, 2004), ClustalO (Sievers et al., 2011) and MAFFT (Katoh et al., 2019) in order to align the Indian 566 SARS-CoV-2 sequences. Subsequently, these aligned results are used to identify the lists of mutations as substitution, deletion, insertion and SNP. Thereafter, a consensus of these results, called as Consensus Multiple Sequence Alignment (CMSA), is prepared to have the final list of mutations so that the advantages of all four alignment techniques can be preserved. It is important to mention that the identification of SNPs can be helpful for the classification of virus strain and accordingly designing of vaccine and defining the dose of the vaccine can be done

\* Corresponding author.

E-mail address: [indrajit@nittrkol.ac.in](mailto:indrajit@nittrkol.ac.in) (I. Saha).<sup>1</sup> Equally contributed.



ScienceDirect

Nimisha Ghosh

## Internet of Things

Volume 13, March 2021, 100068

Research article

## IoT-based freezing of gait detection using grey relational analysis

Nimisha Ghosh <sup>a</sup>✉, Indrajit Banerjee <sup>b</sup>✉

Show more ▾

Share Cite

<https://doi.org/10.1016/j.iot.2019.100068>[Get rights and content](#)

## Abstract

With the massive progress of Internet of Things based technologies, the importance of remote monitoring has increased by many folds especially for the assistance of patients suffering from life threatening diseases. This work deals with the detection of Freezing of Gait in patients having Parkinson's disease with the help of wearable acceleration sensors placed on their legs and hips which consequently results in their real-time remote monitoring. This work proposes a FoG detection method based on Grey Relational Analysis which uses the readings from the sensors to predict the presence or absence of freezing in a patient. Additionally, an ensemble learning approach that uses Grey Relational Analysis as the base classification model

RESEARCH ARTICLE

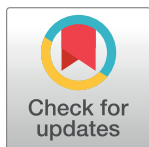
# Phylogenetic analysis of 17271 Indian SARS-CoV-2 genomes to identify temporal and spatial hotspot mutations

Nimisha Ghosh<sup>1,2</sup>, Suman Nandi<sup>3</sup>, Indrajit Saha<sup>3</sup>\*

**1** Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland, **2** Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India, **3** Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

\* These authors contributed equally to this work.

\* [indrajit@nittrkol.ac.in](mailto:indrajit@nittrkol.ac.in)



## OPEN ACCESS

**Citation:** Ghosh N, Nandi S, Saha I (2022) Phylogenetic analysis of 17271 Indian SARS-CoV-2 genomes to identify temporal and spatial hotspot mutations. PLoS ONE 17(3): e0265579. <https://doi.org/10.1371/journal.pone.0265579>

**Editor:** Chandrabose Selvaraj, Alagappa University, INDIA

**Received:** May 27, 2021

**Accepted:** March 7, 2022

**Published:** March 28, 2022

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0265579>

**Copyright:** © 2022 Ghosh et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are available on Figshare: <https://doi.org/10.6084/m9.figshare.19328735.v1>.

**Funding:** This work was carried out during the tenure of an ERCIM 'Alain Bensoussan' Fellowship

## Abstract

The second wave of SARS-CoV-2 has hit India hard and though the vaccination drive has started, moderate number of COVID affected patients is still present in the country, thereby leading to the analysis of the evolving virus strains. In this regard, multiple sequence alignment of 17271 Indian SARS-CoV-2 sequences is performed using MAFFT followed by their phylogenetic analysis using Nextstrain. Subsequently, mutation points as SNPs are identified by Nextstrain. Thereafter, from the aligned sequences temporal and spatial analysis are carried out to identify top 10 hotspot mutations in the coding regions based on entropy. Finally, to judge the functional characteristics of all the non-synonymous hotspot mutations, their changes in proteins are evaluated as biological functions considering the sequences by using PolyPhen-2 while I-Mutant 2.0 evaluates their structural stability. For both temporal and spatial analysis, there are 21 non-synonymous hotspot mutations which are unstable and damaging.

Nimisha Ghosh

## Introduction

It is now close to two years since the emergence of SARS-CoV-2, the virus behind the deadly COVID-19 disease and the scientific community is still struggling to put an end to this pandemic. Though India was able to contain the spread in the first wave, the second wave put the entire system in turmoil. In September 2021, around 30,000 <https://www.covid19india.org/> cases were being registered on a daily basis while in the month of May, this figure surpassed 300,000. Scientists and researchers had attributed this surge due to the evolution of this contagious virus which has resulted in Delta (B.1.617.2) variant. Though the vaccination drive in India is in full swing, doubts regarding the efficacy of the vaccine against such mutations cannot be undermined. Apart from Delta, other variants of concern as declared by W.H.O making their rounds are Alpha (B.1.1.7) [1], Beta (B.1.351) [2] and Gamma (P.1) [3] variants. All these



Nimisha Ghosh

# Identification of miRNA Biomarkers for Diverse Cancer Types Using Statistical Learning Methods at the Whole-Genome Scale

Jnanendra Prasad Sarkar<sup>1,2†</sup>, Indrajit Saha<sup>3\*†</sup>, Adrian Lancucki<sup>4†</sup>, Nimisha Ghosh<sup>5</sup>, Michal Wlasnowolski<sup>6</sup>, Grzegorz Bokota<sup>7,8</sup>, Ashmita Dey<sup>2</sup>, Piotr Lipinski<sup>4</sup> and Dariusz Plewczynski<sup>6,8\*</sup>

## OPEN ACCESS

### Edited by:

Natalia Polouliakh,  
Sony Computer Science Laboratories,  
Japan

### Reviewed by:

Jaromir Gumulec,  
Masaryk University, Czechia  
Tao Huang,  
Shanghai Institute for Biological  
Sciences (CAS), China

### \*Correspondence:

Indrajit Saha  
indrajit@nittrkol.ac.in  
Dariusz Plewczynski  
d.plewczynski@cent.uw.edu.pl

<sup>†</sup> These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Systems Biology,  
a section of the journal  
Frontiers in Genetics

**Received:** 25 October 2019

**Accepted:** 03 August 2019

**Published:** 13 November 2020

### Citation:

Sarkar JP, Saha I, Lancucki A,  
Ghosh N, Wlasnowolski M, Bokota G,  
Dey A, Lipinski P and Plewczynski D  
(2020) Identification of miRNA  
Biomarkers for Diverse Cancer Types  
Using Statistical Learning Methods at  
the Whole-Genome Scale.  
*Front. Genet.* 11:982.  
doi: 10.3389/fgene.2020.00982

<sup>1</sup> Data, Analytics & AI, Larsen & Toubro Infotech Ltd., Pune, India, <sup>2</sup> Department of Computer Science & Engineering, Jadavpur University, Kolkata, India, <sup>3</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, India, <sup>4</sup> Computational Intelligence Research Group, Institute of Computer Science, University of Wrocław, Wrocław, Poland, <sup>5</sup> Department of Computer Science and Information Technology, SOA University, Bhubaneswar, India, <sup>6</sup> Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland, <sup>7</sup> Institute of Informatics, University of Warsaw, Warsaw, Poland, <sup>8</sup> Centre of New Technologies, University of Warsaw, Warsaw, Poland

Genome-wide analysis of miRNA molecules can reveal important information for understanding the biology of cancer. Typically, miRNAs are used as features in statistical learning methods in order to train learning models to predict cancer. This motivates us to propose a method that integrates clustering and classification techniques for diverse cancer types with survival analysis via regression to identify miRNAs that can potentially play a crucial role in the prediction of different types of tumors. Our method has two parts. The first part is a feature selection procedure, called the stochastic covariance evolutionary strategy with forward selection (SCES-FS), which is developed by integrating stochastic neighbor embedding (SNE), the covariance matrix adaptation evolutionary strategy (CMA-ES), and classifiers, with the primary objective of selecting biomarkers. SNE is used to reorder the features by performing an implicit clustering with highly correlated neighboring features. A subset of features is selected heuristically to perform multi-class classification for diverse cancer types. In the second part of our method, the most important features identified in the first part are used to perform survival analysis via Cox regression, primarily to examine the effectiveness of the selected features. For this purpose, we have analyzed next generation sequencing data from The Cancer Genome Atlas in form of miRNA expression of 1,707 samples of 10 different cancer types and 333 normal samples. The SCES-FS method is compared with well-known feature selection methods and it is found to perform better in multi-class classification for the 17 selected miRNAs, achieving an accuracy of 96%. Moreover, the biological significance of the selected miRNAs is demonstrated with the help of network analysis, expression analysis using hierarchical clustering, KEGG pathway analysis, GO enrichment analysis, and



[Published: 02 December 2017](#)

# On-demand fuzzy clustering and ant-colony optimisation based mobile data collection in wireless sensor network

[Nimisha Ghosh](#) , [Indrajit Banerjee](#) & [R. Simon Sherratt](#)

[Wireless Networks](#) **25**, 1829–1845 (2019)

**686** Accesses | **20** Citations | [Metrics](#)

Nimisha Ghosh

## Abstract

In a wireless sensor network (WSN), sensor nodes collect data from the environment and transfer this data to an end user through multi-hop communication. This results in high energy dissipation of the devices. Thus, balancing of energy consumption is a major concern in such kind of network. Appropriate cluster head (CH) selection



# Palindromic target site identification in SARS-CoV-2, MERS-CoV and SARS-CoV-1 by adopting CRISPR-Cas technique

Nimisha Ghosh<sup>a,b,1</sup>, Indrajit Saha<sup>c,\*</sup>, Nikhil Sharma<sup>d</sup>

<sup>a</sup> Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland

<sup>b</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India

<sup>c</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India

<sup>d</sup> Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India

## ARTICLE INFO

Edited by Xavier Carette

### Keywords:

CRISPR-Cas  
Primer  
SARS-CoV-2  
Palindrome  
Protospacer Adjacent Motif  
Target site identification

## ABSTRACT

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) associated Cas protein (CRISPR-Cas) has turned out to be a very important tool for the rapid detection of viruses. This can be used for the identification of the target site in a virus by identifying a 3–6 nt length Protospacer Adjacent Motif (PAM) adjacent to the potential target site, thus motivating us to adopt CRISPR-Cas technique to identify SARS-CoV-2 as well as other members of Coronaviridae family. In this regard, we have developed a fast and effective method using *k*-mer technique in order to identify the PAM by scanning the whole genome of the respective virus. Subsequently, palindromic sequences adjacent to the PAM locations are identified as the potential target sites. Palindromes are considered in this work as they are known to identify viruses. Once all the palindrome-PAM combinations are identified, PAMs specific for the RNA-guided DNA Cas9/Cas12 endonuclease are identified to bind and cut the target sites. In this regard, PAMs such as 5'-TGG-3' and 5'-TTTA-3' in NSP3 and Exon for SARS-CoV-2, 5'-GGG-3' and 5'-TGG-3' in Exon and NSP2 for MERS-CoV and 5'-AGG-3' and 5'-TTTG-3' in Helicase and NSP3 respectively for SARS-CoV-1 are identified corresponding to SpCas9 and FnCas12a endonucleases. Finally, to recognise the target sites of Coronaviridae family as cleaved by SpCas9 and FnCas12a, complements of the palindromic target regions are designed as primers or guide RNA (gRNA). Therefore, such complementary gRNAs along with respective Cas proteins can be considered in assays for the identification of SARS-CoV-2, MERS-CoV and SARS-CoV-1.

## 1. Introduction

COVID-19, the disease caused by Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) has affected a lot of people around the globe and has claimed more than 5.3 million lives as of 15<sup>th</sup> December 2021.<sup>2</sup> SARS-CoV-2 belongs to the family of Coronaviridae which also accommodates MERS-CoV and SARS-CoV-1 viruses (Zhou et al., 2020). The symptoms of COVID-19 include cough, fever, dyspnoea, diarrhoea, myalgia (Hosseini et al., 2020) and in some extreme cases may also lead to severe respiratory distress leading to eventual death. Moreover, comorbidity issue in COVID-19 is relatively high and targets different organs like kidney, liver, heart, brain, etc. (Dey et al.,

2020; Qi et al., 2020).

Since its spread, symptom-based diagnosis of COVID-19 is being performed which includes chest X-ray and CT scan, quantitative reverse transcription polymerase chain reaction (qt-PCR) and antibody test. Most recently, another rapid detection method based on CRISPR-Cas has been proposed by the researchers (Broughton et al., 2020b). Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) associated Cas protein (CRISPR-Cas) is an adaptive immune system in prokaryotic organism that can provide resistance to foreign elements. This system has been exploited in recent times as a powerful gene editing tool and for the diagnosis and inactivation of viruses (Jia et al., 2020). The efficiency of CRISPR-Cas technique is dependent on the design of guide RNA

**Abbreviations:** CRISPR, Clustered Regularly Interspaced Short Palindromic Repeats; PAM, Protospacer Adjacent Motif; gRNA, guide RNA.

\* Corresponding author.

E-mail address: [indrajit@nitttrkol.ac.in](mailto:indrajit@nitttrkol.ac.in) (I. Saha).

<sup>1</sup> Equally contributed.

<sup>2</sup> <https://www.worldometers.info/coronavirus/>.

<https://doi.org/10.1016/j.gene.2021.146136>

Received 1 October 2021; Received in revised form 16 December 2021; Accepted 20 December 2021

Available online 6 January 2022

0378-1119/© 2022 Elsevier B.V. All rights reserved.

Nimisha Ghosh



# Strategies for COVID-19 Epidemiological Surveillance in India: Overall Policies Till June 2021

Nimisha Ghosh

Nimisha Ghosh<sup>1†</sup>, Indrajit Saha<sup>2\*†</sup>, Jnanendra Prasad Sarkar<sup>3</sup> and Ujjwal Maulik<sup>4</sup>

<sup>1</sup> Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, India, <sup>2</sup> Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, India, <sup>3</sup> Larsen & Toubro Infotech Ltd., Pune, India, <sup>4</sup> Department of Computer Science and Engineering, Jadavpur University, Kolkata, India

## OPEN ACCESS

### Edited by:

Shisan Bao,  
The University of Sydney, Australia

### Reviewed by:

Zhongqing Xu,  
Shanghai Jiao Tong University School  
of Medicine, China  
Severino Jefferson Ribeiro Da Silva,  
Fiocruz Pernambuco, Brazil

### \*Correspondence:

Indrajit Saha  
indrajit@nittrkol.ac.in

<sup>†</sup>These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Infectious Diseases - Surveillance,  
Prevention and Treatment,  
a section of the journal  
Frontiers in Public Health

**Received:** 11 May 2021

**Accepted:** 21 June 2021

**Published:** 22 July 2021

### Citation:

Ghosh N, Saha I, Sarkar JP and  
Maulik U (2021) Strategies for  
COVID-19 Epidemiological  
Surveillance in India: Overall Policies  
Till June 2021.  
Front. Public Health 9:708224.  
doi: 10.3389/fpubh.2021.708224

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has gripped the entire world, almost paralysing the human race in its entirety. The virus rapidly transmits *via* human-to-human medium resulting in a massive increase of patients with COVID-19. In order to curb the spread of the disease, an immediate action of complete lockdown was implemented across the globe. India with a population of over 1.3 billion was not an exception and took the challenge to execute phase-wise lockdown, unlock and partial lockdown activities. In this study, we intend to summarise these different phases that the Government of India (GoI) imposed to fight against SARS-CoV-2 so that it can act as a reference guideline to help controlling future waves of COVID-19 and similar pandemic situations in India.

**Keywords:** coronavirus, COVID-19, lockdown, pandemic, SARS-CoV-2, unlock, partial lockdown

## 1. INTRODUCTION

Till now, coronavirus disease 2019 (COVID-19), the disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has resulted in more than 177 million people being affected and over 3 million deaths as reported in 'worldometers.info'<sup>1</sup>. In the initial phases of the pandemic, neither any vaccine nor any therapeutic drugs were at the disposal of the human population, thereby making prevention the only way to curb the spread of the virus. Despite active research studies (1–7) in developing preventive vaccine which started quite soon, immediate action was necessary to stop the rapid spread of the virus *via* human-to-human transmission (8, 9), thereby making lockdown the only viable option around the globe; India being no exception. However, imposing lockdown in India, which has a population of over 1.3 billion, is no mean task. Due to the various challenges of implementing lockdown in India, Government of India (GoI) adopted a strategy for COVID-19 epidemiological surveillance by imposing phase-wise lockdown and unlock, which has been found to be a standout example to curb the outbreak. There are studies that have shown the effect of strategies to control the COVID-19 pandemic in the dental field (10), in hospital (11) and in long-term care facilities (LTCFs) (12), by reducing social mixing (13), pertaining to the laboratory-confirmed interval (14) and diagnosis and treatment of patients with cancer (15) during the pandemic. Furthermore, pharmacological (16) and behavioural strategies (17) for the prevention of COVID-19 and the prevention of the same through contact tracing, screening, quarantine, and isolation (18) have also been studied in several research studies. However, there is a distinct lack of

<sup>1</sup><https://www.worldometers.info/coronavirus/>