ERCIM "ALAIN BENSOUSSAN"
FELLOWSHIP PROGRAMME

ERCIM
European Research Consortium
for Informatics and Mathematics

Scientific Report

First name / Family name	Nimisha Ghosh
Nationality	Indian
Name of the *Host Organisation*	University of Warsaw, Poland
First Name / family name of the *Scientific Coordinator*	Prof. Ania Gambin
Period of the fellowship	01/08/2021 to 31/07/2022

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

During the period of ERCIM fellowship, I worked at Faculty of Mathematics, Informatics and Mechanics, University of Warsaw (MIMUW) on a research project that concerns different aspects of SARS-CoV-2 virus encompassing mutations in the virus, phylogenetic analysis, virus identification, protein-protein interactions between the virus and human proteins etc. This project has been carried out under Prof. Ania Gambin. I have provided one research talk and have attended several others conducted by Computational Biology and Bioinformatics group at MIMUW. I have also participated in the ERCIM community event and gave a poster presentation on my works. Moreover, as a part of the research exchange program (REP), I had visited CNR-IASI Rome, Italy to work with Dr. Daniele Santoni. The main research activities followed during my ERCIM Fellowship are summarised below:

Under the fellowship, I have published three works ((1) Phylogenetic analysis of 17271 Indian SARS-CoV-2 genomes to identify temporal and spatial hotspot mutations, (2) A review on evolution of emerging SARS-CoV-2 variants based on spike glycoprotein, (3) Palindromic target site identification in SARS-CoV-2, MERS-CoV and SARS-CoV-1 by adopting CRISPR-Cas technique) and two have been submitted for evaluation while one work is still ongoing.

The first work considers multiple sequence alignment (MSA) of 17271 Indian SARS-

CoV-2 genomes using multiple alignment using fast fourier transform (MAFFT) followed by their phylogenetic analysis using Nextstrain to eventually identify hotspot mutations both month-wise (temporal) and state-wise (spatial). 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, thereby resulting in 130 and 250 hotspot mutations respectively. 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. The hotspot mutations which are unstable and damaging and common in both the categories are T77A and V149A in NSP6, T95I and E484Q in Spike, Q57H and T223I in ORF3a, I82S and I82T in Membrane, D119V and F120L in ORF8, R203K, R203M and G215C in Nucleocapsid. Furthermore, as recognised by virologists, E484K in Spike which is identified in temporal analysis is yet another major mutation which is responsible for improving the ability of the virus to escape the host's immune system.

In the second work, multiple sequence alignment of 77681 SARS-CoV-2 genomes of 98 countries over the period from January 2020 to July 2021 has been performed using MAFFT followed by phylogenetic analysis to analyse the mutations in Spike glycoprotein. 12 different important variants identified so far are Alpha, Beta, Eta, Epsilon, Iota, Kappa, Delta, Lambda, Gamma, Zeta, Theta and Omicron. These variants have 84 unique mutations and include some notable mutations like K417N, L452R, S477N, T478K, E484K/Q, N501Y, D614G, P681H/R, Y144-, H69- and V70-. Furthermore, the characteristics of the variants are elaborately discussed along with their specific mutations. Thereafter, the individual evolution of these mutation points are visualised along with their evolution in the respective variants. Moreover, the characteristics of the non-synonymous mutation points (substitutions) are judged by evaluating their biological functions by considering the sequences and using PolyPhen-2 while I-Mutant 2.0 evaluates the protein structural stability. Thus, this work provides a comprehensive review of the emerging variants and the characteristics of the corresponding mutation points along with the effects of vaccine and therapeutics on the variants.

The third work has adopted the concept of CRISPR-Cas system to identify the target sites for the identification of SARS-CoV-2 and other viruses of Coronaviridae family, that is MERS-CoV and SARS-CoV-1. In this regard, identification of protospacer adjacent motif or PAM is carried out in this work. PAM is a short DNA sequence having usually a length of about 3–6 nt that is present adjacent to CRISPR in the genomic sequence. The genomic locations that are the potential target sites for the identification of viruses are limited by the presence and locations of the PAM. Thus, in order to find the target sites for the identification of SARS-CoV-2, MERS-CoV and SARS-CoV-1 viruses, initially the PAM and their corresponding genomic locations are identified. Once the PAM are identified, instead of finding short palindromic repeats as required by CRISPR-Cas, we have modified the idea to consider palindromic sequences which are adjacent to PAM to be the target sites for virus identification. Thereafter, to bind and cut the target sites, specific PAMs are identified for the RNA-guided DNA Cas9/Cas12 endonuclease. 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. It is worth mentioning that studies performed by Cain et al., 2001, Dirac et al., 2002, Chew et al., 2004 have suggested that palindromes can be considered to be involved in target identification, viral packaging and defence mechanisms. A palindromic sequence is a symmetrical sequence so that when read from the reverse direction, it is the exact complement of itself. For example, TGCA is a palindrome of length 4. It is to be noted that a palindrome is always even in length. Thereafter, to recognise these target sites in a virus genome as cleaved by SpCas9 and FnCas12a, primers are designed as complementary to the target site sequences. Thus, these complementary palindromic primers can be considered in assays for the rapid identification of SARS-CoV-2, MERS-CoV and SARS-CoV-1. These primers are akin to guide RNA (gRNA) in CRISPR-Cas technology.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

During the ERCIM training program at MIMUW University of Warsaw, I have worked on six research papers in which three papers are published and two papers are submitted for publication while one work with Prof. Ania Gambin is currently ongoing. List of published papers are:

1. **N Ghosh, S Nandi, I Saha, Phylogenetic analysis of 17271 Indian SARS-CoV-2 genomes to identify temporal and spatial hotspot mutations.** Published in PloS one, Volume 17, Pages e0265579, 2022.

   [**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.]

2. **N Ghosh, S Nandi, I Saha, A review on evolution of emerging SARS-CoV-2 variants based on spike glycoprotein.** Published in International Immunopharmacology, Volume 105, Pages 108565, 2022.

   [**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.]

3. **N Ghosh, I Saha, N. Sharma, Palindromic Target Site Identification in SARS-CoV-2, MERS-CoV and SARS-CoV-1 by Adopting CRISPR-Cas Technique.** Published in Gene, Volume 818, Pages 146136, 2022.

   [**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.]

List of submitted papers are:

D. Santoni, N. Ghosh, C. Derelitto, I. Saha, Study the effects of Transcription Factors on Human Proteins interacting with Spike Glycoprotein of SARS-CoV-2. Submitted to Infection, Genetics and Evolution.

N. Ghosh, I. Saha, Unveiling the Biomarkers of Cancer and COVID-19 and Their Regulations in Different Organs by Integrating RNA-Seq Expression and Protein-Protein Interactions. Submitted to Journal of Translational Medicine.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

During my tenure as an ERCIM fellow, I have presented my work in several seminars and have also attended a workshop on COVID-19. The details of the seminars and the workshop are as follows:

Seminars:

1) Delivered a virtual talk on computational biology and my work at Techno International New Town, Kolkata, India on 15th September 2021.
2) Presented a poster pertaining to my research work on 9th November 2021 at ERCIM community event.
3) Delivered a talk on "Interactome based Machine Learning predicts Potential Therapeutics for COVID-19" at Faculty of Mathematics, Informatics and Mechanics (Computational Biology and Bioinformatics group), University of Warsaw, Warsaw, Poland on 2nd March 2022.

Workshop:

1) Virtually attended "Taxila: Empowering the fight against COVID-19 through text" workshop arranged jointly by University of Warsaw and The Systems Biology Institute, Japan

IV – RESEARCH EXCHANGE PROGRAMME (REP)

Duration: 21st March 2022-25th March 2022
Place: CNR-IASI, Rome, Italy,
Scientific Contact: Dr. Daniele Santoni

During my visit to CNR-IASI, Dr. Santoni and I worked on common research project pertaining to COVID-19. In this regard, we have worked on the effects of transcription factors on human proteins interacting with Spike Glycoprotein of SARS-CoV-2. The visit was quite fruitful and I learnt quite a lot of things from Dr. Santoni. As a consequence of this visit, we have successfully completed one work and submitted the same to Infection, Genetics and Evolution journal for further consideration. We have also planned to collaborate on other projects in the future.

Anna Gambin; Uniwersytet Warszawski
Digitally signed by Anna Gambin; Uniwersytet Warszawski
Date: 2022.05.25 09:46:24 +02'00'

Nimisha Ghosh

Name and Signature of Fellow (Nimisha Ghosh)

Name and Signature of Scientific Coordinator (Prof. Ania Gambin)