Volume 64, Issue 3, 2020

**Journal of Scientific Research**

Institute of Science,

Banaras Hindu University, Varanasi, India.

National Conference on Frontiers in Biotechnology & Bioengineering (NCFBB 2020), JNTU Hyderabad, India

# *Insilico* Epitope Mapping of SARS-CoV-2

A. Shashanka[*1], A.Uma[1], B.Suresh Babu[1]

[1]Centre for Biotechnology, Institute of Science and Technology, JNTUH, Hyderabad, India.
shashi.arendra@gmail.com, vedavathi1@jntuh.ac.in, bastipati@gmail.com

***Abstract:*** **Severe Acute Respiratory Syndrome CoronaVirus-2 (SARS CoV-2) causes COVID-19, an ongoing pandemic. In the world, till today more than one crore people got infected. In this fight against the COVID-19, there is an urgent need for the development of better diagnosis, therapy, and vaccines. For this development, Antigen-Antibody interactions are to be studied which is a chemical reaction between antibodies produced by B-cells of white blood cells and antigens during the immune response. This binding takes place between the epitope of the antigen and the paratope of an antibody. Identifying these crucial epitopes of the virus can help in the development of better serum antibodies detecting kits and therapies involving antibodies. Here, we have identified a few epitopes from a genome of the SARS-COV 2 virus confined from an Indian patient using different bioinformatic tools and ranked them based on properties like surface accessibility, hydrophilicity, and antigenicity.**

***Index Terms:*****SARS CoV-2, COVID19, antigen-antibody interaction, epitopes, *in silico*, surface accessibility, hydrophilicity, antigenicity**

## I. INTRODUCTION

Coronavirus is the positive-sense single-stranded RNA virus that is prevalent all over the world. The virus is transmitted through close contact between people, often through small droplets generated by coughing, sneezing, and talking. The droplets do not travel long distances but fall on the ground or onto the surface. When people get in contact with the contaminated surface and touch their face,the infection passes on to them. Some people are asymptomatic from whom the spread is possible. The clinical manifestations shown are respiratory illness, fever, anosmia, dry cough, sore throat, body pains, itchy skin, discoloration of fingers, etc. The term 'coronavirus' was constituted by June Almeida and David Tyrrells from the Latin word which means 'crown' because the structure contains bulbous-like projections on its surface. The morphology is due to viral spike peplomers, which are viral surface proteins. There are four types of coronavirus. They are alpha, beta, delta, and gamma. The first two genera infect humans while the latter two genera infect birds. The known viruses that infect humans are:

1. Human coronavirus 229E (HCoV-229E)
2. Human coronavirus NL63 (HCoV-NL63)
3. Human coronavirus OC43 (HCoV-OC43)
4. Human coronavirus HKU1 (HCoV-HKU1)
5. Middle East Respiratory Syndrome-related coronavirus (MERS-CoV)
6. Severe Acute Respiratory Syndrome coronavirus (SARS-CoV)
7. Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2)

SARS-CoV-2 is the pandemic that broke out in late 2019 in Wuhan city of China. Within a few months, it transmitted across the globe affecting more than 190 nations. As mentioned above, coronavirus has positive-sense single-stranded RNA. The genome length ranges from 26 to 32 kilobases. Out of this, two-third of the genome encodes for 16 non-structural proteins while one-third genome encodes for structural and accessory proteins. 16 non-structural proteins play a role in viral replication, proteolytic maturation of proteins, and stability of viruses from host immune responses. The structural proteins incorporate Spike protein (S), Membrane protein (M), an Envelope protein (E), RNA bound Nucleocapsid protein (N), and Hemagglutinin Esterase (H) which is present only in beta coronavirus. The Spike protein is composed of S1 and S2 subunit which is a class I fusion protein that intervenes the receptor binding and membrane fusion between host cells and viruses. The S1 subunit creates the head of the spike and receptor-binding domain (RBD) while the S2 subunit binds the stem that harbors the spike in the viral envelope and on protease activation permits fusion. The E protein forms the viral envelope and the M protein maintains its structural shape. The N protein is present inside the

*Corresponding Author

*DOI:* 10.37398/JSR.2020.640311
Selection and Peer-review under responsibility of the Program Chairs.

viral envelope which is encircled to a positive-sense single-stranded RNA genome in a continuous beads-on-a-string type conformation. When the virus is outside the host cell, lipid bilayer envelope, membrane protein, and nucleocapsid are accountable for virus salvation.

*A. Antigen-Antibody interactions*

The chemical reaction between antigens and antibodies formed by B-cells of white blood cells at the time of an immune response is called Antigen-Antibody Interactions. This reaction protects the body from complex foreign molecules like pathogens and their chemical toxins. The antigen with high specificity and affinity binds with antibodies forming an antigen-antibody complex. This complex is deactivated or destroyed in cellular systems. The antigen binds to the paratope of an antibody with the help of its epitope. This binding occurs by weak and non-covalent interactions like electrostatic interactions, hydrogen bonds, VanderWaals forces, and hydrophobic interactions. These interactions help in the clinical diagnosis of Syphilis, HIV, Rubella, and in the ABO blood grouping.

*B. Epitope Mapping*

Epitope mapping is the technique of recognizing the binding site or epitope of an antibody on its target antigen. This identification and description of these epitopes help in the development of serum antibody detecting kits, antibody-dependent therapies. Scientists all over the world are making their efforts to initially sequence the genome of the virus. Now most of them are working on the progressing of diagnostic kits and vaccine development. With the small effort, epitope mapping is studied and performed on the genomic sequence of the Indian sample.

*C. In Silico Source & Tools Used*

To identify the epitopes from the gained genome sequence various bioinformatics sources & tools have been used. They are:

- GISAID (Global Initiative Serving All Influenza Data)
- GeneMark
- SOSUI
- IEDB Analysis Resource
- Peptide 2.0

***GISAID*** is the primary source for genomic data that is a global science initiative. This holds the genomic sequences of influenza viruses till novel coronavirus that is liable for COVID-19. From the day of its inauguration in 2008, it is a substitute for sharing avian influenza data including the outbreak of the data of the H1N1 pandemic in 2009, the H7N9 epidemic in 2013, and the COVID19 pandemic in 2020. It made a great effort for COVID19 causing the SARS COV-2 virus by understanding the genome sequences modeled in real-time. This helps in detecting viral mutations over the world. Till June 2020, around 54,000 SARS COV-2 genome sequences are updated by more than 450 laboratories all over the world in the GISAID database.

***GeneMark*** developed at Georgia Institute of Technology in Atlanta is the gene prediction program. It was developed in 1993 while used initially in 1995 as a primary gene prediction tool for the bacterial genomic sequence of *Haemophilus influenza*, and in 1996 aimed for the archaeal genome of *Methanococcusjannaschii*. Its algorithm brought in inhomogeneous three Markov chain models for protein-coding DNA sequence and the Bayesian model for gene prediction in double-stranded DNA.

Table I.Different tools used for analysis

| S.No. | Family | Program |
|---|---|---|
| 1 | Bacteria, Archaea | • GeneMark<br>• GeneMarkS<br>• GeneMarkS+ |
| 2 | Metagenomes and Metatranscriptomes | • MetaGeneMark |
| 3 | Eukaryotes | • GeneMark<br>• GeneMark.hmm<br>• GeneMark-ES<br>• GeneMark-ET<br>• GeneMark-EX |
| 4 | Viruses, phages, and plasmids | • Heuristic models |
| 5 | Transcripts assembled from RNA-Seq read | • GeneMarkS-T |

***SOSUI*** is an online tool that helps in indicating the secondary structure of proteins from a given amino acid sequence. This innovation was developed by TOKYO University which means hydrophobic. The ambition is to figure out if a given protein sequence is a soluble or transmembrane protein.

The solubility of amino acids is predicted by SOSUI considering 4 features. They are:

- Hydropathy Index
- Amino acid charge
- Amphilicity index
- Length of amino acid sequence

The results show the hydropathy index, helical wheel diagram of transmembrane protein including their length, length of an amino acid sequence.

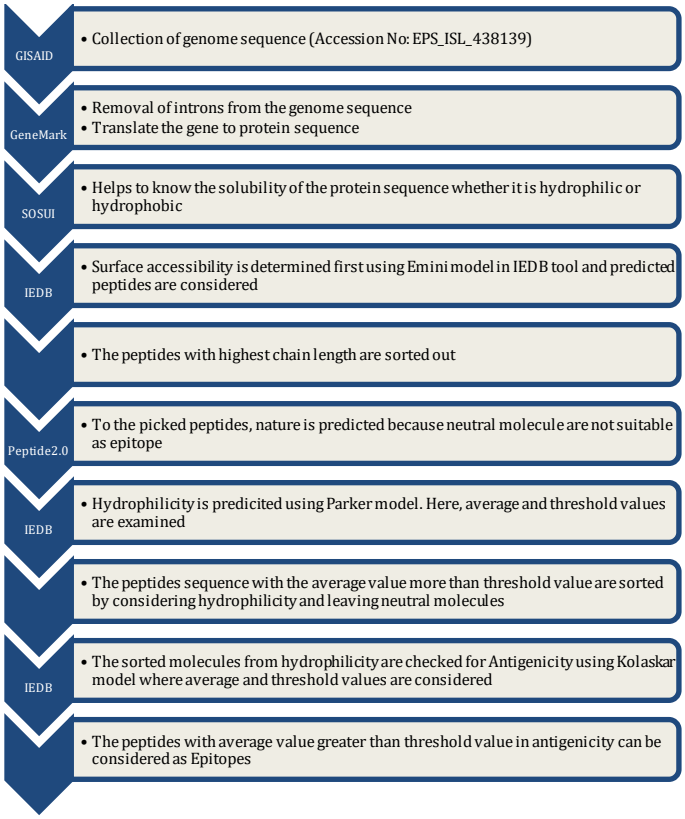
***IEDB Analysis Resource*** came up with several tools that are useful for the prophecy and evaluation of immune epitopes. The tools that fall into this category are:

- T Cell Epitope Prediction Tools – MHC I & II binding predictions, peptide processing prediction, and immunogenicity predictions
- B Cell Epitope Prediction Tools – predict regions of proteins that are likely to be epitopes in B-cell response
- Analysis Tools – an elaborate study of a known sequence of epitope or a group of sequences

B cell epitope prediction tools are ElliPro, Discotope, and Prediction of linear epitopes from protein sequence. Along with the prediction of the epitopes, IEDB Analysis Resources provides the various properties to study the epitopes' characteristics. The properties are surface accessibility, flexibility, Beta-turn, antigenicity, hydrophilicity, etc. In this tool, to figure the surface accessibility Emini method is used, and to study the hydrophilicity Parker method is used, and to detect the antigenicity Kolaskar and Tongaonkar model is used.

***Peptide 2.0*** is a company to synthesize the second generation peptide. It provides peptides to customers by taking orders. It contains peptide tools like Peptide hydrophobicity/hydrophilicity analysis, Peptide property calculator, Peptide molecular weight calculator, Peptide generator, etc. These tools are useful to know the properties like an isoelectric point, pH, pKa, molecular weight, and many more characteristics of the sequences that can be determined by using this tool.

## II. METHODOLOGY




## III. RESULTS AND DISCUSSION

Epitopes vary based on the considered predicted sequences that are either by combining continuous adjacent peptide sequence or directly using predicted peptides. A total of 111 epitopes were obtained by combining continuous adjacent sequences of 10mer with the help of the IEDB Analysis & Resources(Leonard A..et al 2020). Here to sort out a limited number of epitopes, various bioinformatics tools are used like GISAID, GeneMark, IEDB Analysis Resource, Peptides 2.0, and SOSUI. From the GISAID, the genome sequence of size 29,904 is selected which is the Hyderabad sample. Then, this genome data of the sample is given as an input in the GeneMark tool to remove introns and translate the gene into protein sequence. After removing introns, the 10 gene sequences remained which are exons. These 10 genes are initially checked for solubility using the SOSUI tool.

As the all 10 genes are soluble sequences, then surface accessibility of the genes are identified using IEDB Analysis Resource. From the predicted peptides of the given genome, the sequences with the maximum chain length are chosen, as their surface of exposure is more. Then, to the obtained peptides their nature is predicted using Peptide 2.0, and hydrophilicity is determined using the IEDB tool, where the average value greater than the threshold value is again sorted out leaving neutral molecules. This is succeeded by the prophecy of the antigenicity of the sorted peptides with the help of the IEDB tool. Here, the threshold and average values are noted initially for all the selected peptides. The sequences whose average value exceeds the threshold value are determined as “EPITOPES”.

By analyzing the tabular data, the recognized epitopes are:

- Gene 2: QAMTQMYKQARSEDKRAK
- Gene 3: YVDTPNNTDFSRVSAKPPPGDQF

: VEWKFYDAQPCSDKAYKIEEL

- Gene 9: AGSKSPIQ

These epitopes might be used for establishing the serological test by predicting the paratope and determining the affinity between them and also used for vaccine designing.

Table II. The table shows the results of the final epitopes from the overall genome gained from GISAID

| S. No. | Gene No. | Peptide sequence | Chain length | Nature | Hydrophilicity | | Antigenicity | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Average | Threshold | Average | Threshold |
| 1 | Gene 1 | HEIAWYTERSEKSYELQT | 18 | acidic | 3.204 | 2.993 | 0.971 | 0.975 |
| | | PKLDNYYKKDNSYFT | 15 | basic | 3.477 | 3.462 | 0.98 | 0.982 |
| 2 | Gene 2 | **QAMTQMYKQARSEDKRAK** | **18** | **basic** | **3.884** | **3.851** | **0.961** | **0.958** |
| 3 | Gene 3 | KWGKARLYYDSMSYEDQDALFA | 22 | acidic | 2.449 | 2.354 | 1.002 | 1.004 |
| | | **YVDTPNNTDFSRVSAKPPPGDQF** | **23** | **Acidic** | **3.396** | **3.592** | **0.998** | **0.991** |
| | | **VEWKFYDAQPCSDKAYKIEEL** | **21** | **Acidic** | **2.944** | **2.7** | **1.041** | **1.035** |
| 4 | Gene 4 | VYDPLQPELDSFKEELDKYFKNHTSPD | 27 | Acidic | 2.347 | 2.318 | 1.002 | 1.005 |
| 5 | Gene 5 | - | - | - | - | - | - | - |
| 6 | Gene 6 | SRTLSYYKLGA | 11 | Basic | 0.94 | 0.454 | 1.046 | 1.07 |
| 7 | Gene 7 | - | - | - | - | - | - | - |
| 8 | Gene 8 | - | - | - | - | - | - | - |
| 9 | Gene 9 | YSKWYIRVGAR | 11 | Basic | 0.175 | -0.846 | 1.036 | 1.047 |
| | | **AGSKSPIQ** | **8** | **Basic** | **3.225** | **3.221** | **1.015** | **1.011** |
| 10 | Gene 10 | AAEASKKPRQKRTAT | 15 | Basic | 4.718 | 4.684 | 0.964 | 0.964 |
| | | TFPPTEPKKDKKKKADETQALPQRQKKQ | 28 | Basic | 4.739 | 4.523 | 0.966 | 0.971 |

## CONCLUSION

For the genome sequence gained from the primary source GISAID, the epitopes are predicted using various bioinformatics tools. These epitopes are applicable for the advancement of antibody-dependent therapies, vaccine designing, and in progressing of diagnostic kits.

## ACKNOWLEDGMENT

We gratefully acknowledge the authors from the originating laboratories responsible for obtaining the specimens and submitting laboratories where genetic sequence data is generated and shared via the GISAID initiative on which research is based.

## REFERENCES

Almeida JD, Berry DM, Cunningham CH, Hamre D, Hofstad MS, Mallucci L, McIntosh K, Tyrrell DA (November 1968). "Virology: Coronaviruses". Nature. 220 (5168):650. Bibcode:1968Natur.220..650.. doi:10.1038/220650b0

Besemer J. and Borodovsky M. "Heuristic approach to deriving models for gene finding." Nucleic Acids Research (1999) 27 (19): 3911–3920. doi:10.1093/nar/27.19.3911

Besemer J. and Borodovsky M. "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses." Nucleic Acids Research (2005) 33 (Web Server Issue): W451-454. doi:10.1093/nar/gki487

Boyle, Alan (June 11, 2020). "Amid COVID-19 pandemic, experts lay out 10-point plan for a genomic revolution in public health". GeekWire. Retrieved June 14, 2020.

Braden, BC; Dall'Acqua, W; Eisenstein, E; Fields, BA; Goldbaum, FA; Malchiodi, EL; Mariuzza, RA; Schwarz, FP; Ysern, X; Poljak, RJ (1995). "Protein motion and lock and key complementarity in antigen-antibody reactions". PharmaceuticaActaHelvetiae. 69 (4): 225–30. doi:10.1016/0031-6865(94)00046-x. PMID 7651966.

Chang CK, Hou MH, Chang CF, Hsiao CD, Huang TH (March 2014). "The SARS coronavirus nucleocapsid protein—forms and functions". Antiviral Research. 103:39-50. doi:10.1016/j.antiviral.2013.12.009. PMC 7113676. PM ID 24418573.

Cherry, James; Demmler-Harrison, Gail J.; Kaplan, Sheldon L.; Steinbach, William J.; Hotez, Peter J. (2017). Feigin and Cherry's Textbook of Pediatric Infectious Diseases. Elsevier Health Sciences. p. PT6615. ISBN 978-0-323-39281-5.

"Definition of Coronavirus by Merriam-Webster". Merriam-Webster. Archived from the original on 2020-03-23. Retrieved 2020-03-24.

DeLisser, HM (1999). "Epitope mapping". Adhesion Protein Protocols. Methods Mol Biol. 96. pp. 11–20. doi:10.1385/1-59259-258-9:11. ISBN 978-1-59259-258-6. PMID 10098119.

Elbe, S., and Buckland-Merrett, G. (2017) Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges, 1:33-46. doi:10.1002/gch2.1018 PMCID: 31565258

Fehr AR, Perlman S (2015). "Coronaviruses: an overview of their replication and pathogenesis". In Maier HJ, Bickerton E, Britton P (eds.). Coronaviruses. Methods in Molecular Biology. 1282. Springer. pp. 1–23. doi:10.1007/978-1-4939-2438-7_1. ISBN 978-1-4939-2438-7. PMC 4369385. PMID 25720466.

Hirokawa T., Boon-Chieng S., and Mitaku S., Bioinformatics, 14 378-9 (1998) SOSUI: classification and secondary structure prediction system for membrane proteins.

Kolaskar, S., &Tongaonkar, P. C. A semi-empirical method for prediction of antigenic determinants on protein antigens. Febbs Lett, 276(1–2), 172–174 (1990).

Leonardo A. Guevarra Jr., Gianne Eduard L. Ulanday (2020) Immune epitope map of reported protein sequence of SARS COV2 doi:10.21203/rs.3.rs-18689/v1

Li F, Li W, Farzan M, Harrison SC (September 2005). "Structure of SARS coronavirus spike receptor-binding domain complexed with receptor". Science. 309 (5742):1864-68. Bibcode:2005Sci...309.1864L. doi:10.1126/science.1116480. PMID 16166518. S2CID 12438123

Martinez-Garcia, et al. (2014). "Unveiling viral-host interactions within the microbial dark matter". Nature communications, 5.

McDowell, Robin (May 15, 2008). "Indonesia hands over bird flu data to new database". Fox News. Associated Press. Retrieved June 7, 2020.

Mitaku S., Hirokawa T. Protein Eng. 11 (1999) Physicochemical factors for discriminating between soluble and membrane proteins: hydrophobicity of helical segments and protein length

Mitaku S., Hirokawa T., and Tsuji T., Bioinformatics, 18 608-16(2002)Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces.

Monto AS, DeJonge P, Callear AP, Bazzi LA, Capriola S, Malosh RE, et al. (April 2020). "Coronavirus occurrence and transmission over 8 years in the HIVE cohort of households in Michigan". The Journal of Infectious Diseases. 222: 9–16. doi:10.1093/infdis/jiaa161. PMC 7184402. PMID 32246136.

Neuman BW, Kiss G, Kunding AH, Bhella D, Baksh MF, Connelly S, et al. (April 2011). "A structural analysis of M protein in coronavirus assembly and morphology". Journal of Structural Biology. 174 (1): 11–22. doi:10.1016/j.jsb.2010.11.021. PMC 4486061. PMID 21130884.

Parker, J. M. R., Guo, D., & Hodges, R. S. New Hydrophilicity Scale Derived from High-Performance Liquid Chromatography Peptide Retention Data: Correlation of Predicted Surface Residues with Antigenicity and X-ray-Derived Accessible Sites. Biochemistry, 25(19), 5425–5432 (1986).

Sela-Culang, Inbal; Kunik, Vered; Ofran, Yanay (2013). "The structural basis of antibody-antigen recognition". Frontiers in Immunology. 4: 302. doi:10.3389/fimmu.2013.00302. PMC 3792396. PMID 24115948.

Shu, Y., McCauley, J. (2017) GISAID: Global initiative on sharing all influenza data – from vision to reality. EuroSurveillance, 22(13) doi:10.2807/1560-7917.ES.2017.22.13.30494 PMCID: PMC5388101

Stein, L. (2001). "Genome annotation: From sequence to biology." Nature Reviews Genetics 2(7): 493-503.

Sturman LS, Holmes KV (1983-01-01). Lauffer MA, Maramorosch K (eds.). "The molecular biology of coronaviruses". Advances in Virus Research. 28: 35–112. doi:10.1016/s0065-3527(08)60721-6. ISBN 9780120398287. PMC 7131312. PMID 6362367.

Tyrrell DA, Fielder M (2002). Cold Wars: The Fight against the Common Cold. Oxford University Press. p. 96. ISBN 978-0-19-263285-2

Thomas, Liji (June 28, 2020). "Genomics used to trace origin of SARS-CoV-2 in Canada". News Medical. Retrieved June 29,2020.

Van Oss, CJ; Good, RJ; Chaudhury, MK (1986). "Nature of the antigen-antibody interaction. Primary and secondary bonds: optimal conditions for association and dissociation". Journal of Chromatography. 376: 111–9. PMID 3711190

Vita, R. et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res, 47(D1), D339–D343. https://doi.org/10.1093/nar/gky1006 (2019)

Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2018 Oct 24. doi: 10.1093/nar/gky1006. [Epub ahead of print] PubMed PMID: 30357391.

Wertheim JO, Chu DK, Peiris JS, Kosakovsky Pond SL, Poon LL(June 2013). "A case for the ancient origin of coronaviruses". Journal of Virology. 87 (12): 7039–45. doi:10.1128/JVI.0327312. PMC 3676139. PMID 23596293.

Westwood, Olwyn M. R.; Hay, Frank C., eds. (2001). Epitope Mapping: A Practical Approach. Oxford, Oxfordshire: Oxford University Press. ISBN 978-0-19-963652-5

Woo PC, Huang Y, Lau SK, Yuen KY (August 2010). "Coronavirus genomics and bioinformatics analysis". Viruses. 2(8):180420. doi:10.3390/v2081803. PMC 3185738PMID 21994708. Coronaviruses possess the largest genomes [26.4 kb (ThCoV HKU12) to 31.7 kb (SW1)] among all known RNA viruses

www.iedb.org

https://www.peptide2.com/

https://www.mygov.in/covid-19/

https://www.slideshare.net/Haddies/antigen-and-antibody-reaction

https://www.slideshare.net/taha244ali/2-antigens-immunogens-epitopes-and-haptens

***