Communications in Information and Systems
Volume 21, Number 1, 31–36, 2021

SARS-CoV-2 becoming more infectious as revealed by algebraic topology and deep learning

Jiahui Chen, Rui Wang, and Guo-Wei Wei*

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) caused by coronavirus disease 2019 (COVID-19) has led to a tremendous human fatality and economic loss. SARS-CoV-2 infectivity is a key reason for the widespread viral transmission, but its rigorous experimental measurement is essentially impossible due to the on-going genome evolution around the world. We show that artificial intelligence (AI) and algebraic topology (AT) offer an accurate and efficient alternative to the experimental determination of viral infectivity. AI and AT analysis indicates that the on-going mutations make SARS-CoV-2 more infectious.

AMS 2000 subject classifications: Primary 68T01, 68U01, 92B05; secondary 00A69.
Keywords and phrases: Viral infectivity, binding affinity change, mutation, deep learning, persistent homology.

The expeditious spread of coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has led to over 40 million confirmed cases and over one million fatalities. In the 21st century, three major outbreaks of deadly pneumonia are caused by β-coronaviruses: SARS-CoV (2002), Middle East respiratory syndrome coronavirus (MERS-CoV) (2012), and SARS-CoV-2 (2019). However, SARS-CoV-2 has an unprecedented scale of infection and potentially becomes a seasonal disease.

SARS-CoV-2 infectivity is a vital factor in COVID-19 transmission, prevention, and economic reopening. Viral infectivity is related to the viral infection rate of a population, which can be affected by the prevention measures, such as social distancing, use of masks, quarantine, contact tracing, etc. The intrinsic viral infectivity can be determined by virus quantification that counts the number of viruses in a specific volume over a unit of time by using either traditional or modern methods [5]. The former includes plaque assay, focus forming assay, endpoint dilution assay, protein assay,

*Corresponding author. E-mail: weig@msu.edu.

hemagglutination assay, bicinchoninic acid assay, single radial immunodiffusion assay, and transmission electron microscopy. The latter has tunable resistive pulse sensing, flow cytometry, quantitative polymerase chain reaction, and enzyme-linked immunosorbent assay (ELISA). Traditional methods are generally slow and labor-intensive, while modern methods can significantly reduce quantification time. Among these methods, ELISA is based on protein-protein interactions (PPIs), such as antibody-antigen binding events being counted by chromogenic or fluorescence reporters. Both traditional and modern methods for viral infectivity measurement are expensive and time-consuming. Epidemiological and biochemical studies show that the infectivity of different SARS-CoV strains in host cells is proportional to the binding free energy between the spike (S) protein receptor-binding domain (RBD) of each strain and angiotensin-converting enzyme 2 (ACE2) expressed by host cells [11]. ACE2 is a single-pass transmembrane protein with its active domain exposed on the cell surface and is expressed in the lungs and many other tissues. It is the main cell entry point for SARS-CoV and SARS-CoV-2, and some other coronaviruses. The cell entry is primed by TMPRSS2 (transmembrane serine protease 2) [5].

For an on-going pandemic, viral infectivity can be further changed by viral evolution [17]. Mutagenesis is a basic biological process that changes the genetic information of organisms, which serves as a primary source for infectivity variation and many kinds of cancers and heritable diseases, as well as a driving force for natural evolution. SARS-CoV-2 belongs to the coronaviridae family and the Nidovirales order, which has been shown to have a genetic proofreading mechanism in its replication achieved by an enzyme called non-structure protein 14 (NSP14) in synergy with NSP12, i.e., RNA-dependent RNA polymerase. Therefore, SARS-CoV-2 has a higher fidelity in its transcription and replication process than that of other single-stranded RNA viruses, such as flu virus and HIV. In general, the frequency of virus mutations is accumulated by the natural selection, cellular environment, polymerase fidelity, random genetic drift, features of recent epidemiology, host immune responses, gene editing, replication mechanism, etc [9, 13]. Although it is difficult to determine the detailed mechanism of a specific mutation, mutations tracked by single nucleotide polymorphism (SNP) calling provide a method to understand the molecular mechanism of SARS-CoV-2 proteins, PPIs, and their synergy with host cell proteins, enzymes, and signaling pathways. By applying SNP calling to more than 60,000 genome isolates deposited at the GISAID database (https://www.gisaid.org/) we found over 18,000 single mutations, mostly caused by host gene editing [13], compared with the first SARS-CoV-2 genome collected on December 24, 2019, in

Wuhan [15]. In our work, we use the Cluster Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) for multiple sequence alignment. We identify 6 SARS-CoV-2 substrains in the world by using the K-means clustering based on pairwise Jaccard distances between sequences [2]. Since there are more than 2,000 unique mutations found for the SARS-CoV-2 S protein gene [13], it is extremely challenging to experimentally determine the virus infectivity associated with various SARS-CoV-2 genetic variants found around the world.

The computational estimation of mutation-induced protein-protein binding free energy (BFE) changes is an important approach for understanding the impact of mutations on PPIs and viral infectivity. A variety of advanced methods has been developed [8], and their performance can be validated by standard databases, including SKEMPI (https://life.bsc.es/pid/skempi2/) for PPIs and AB-Bind, a database for mutation-induced antibody-antigen complex BFE changes [10]. However, due to the challenges of the intricate complexity of PPIs and relatively scarce experimental data, limited success has been achieved for the AB-Bind problem.

Mathematical techniques, such as persistent homology [1, 4, 7, 16], evolutionary de Rham-Hodge theory [3], and persistent spectral graph [14] provide an essential representation with a controllable simplification of biomolecular complexity and retain crucial physical and biological information of PPIs. A deep learning algorithm called NetTree, which combines convolutional neural networks and gradient boosting trees, was constructed to tackle the challenge of scarce data. The resulting method that integrates the topological representation of complex data and NetTree, called TopNetTree, was about 22% better than the previous best result for the AB-Bind dataset and significantly outperformed the state-of-the-art in the literature on the SKEMPI database [12]. Figure 1 illustrates the architecture of the TopNetTree. TopNetTree was retrained on a large dataset of 8338 PPI entries to improve its reliability before applied to predict the BFE changes following mutations on the S protein RBD [2]. By examining the mutation-induced BFE changes of the ACE2 and S protein complex, it was found that three out of six SARS-CoV-2 substrains have become slightly more infectious, while the other three substrains have significantly strengthened their infectivity [2]. We also found that SARS-CoV-2 is slightly more infectious than SARS-CoV [2]. A mutation at the residue 614 of the S protein was also reported to make SARS-CoV-2 infectious [6].

It is imperative to have geographic- and demographic-specific strategies in virus control, containment, prevention, and medication. Such strategies depend on our understanding of how mutations have changed the SARS-CoV-2 structure, function, activity, infectivity, and virulence of various viral

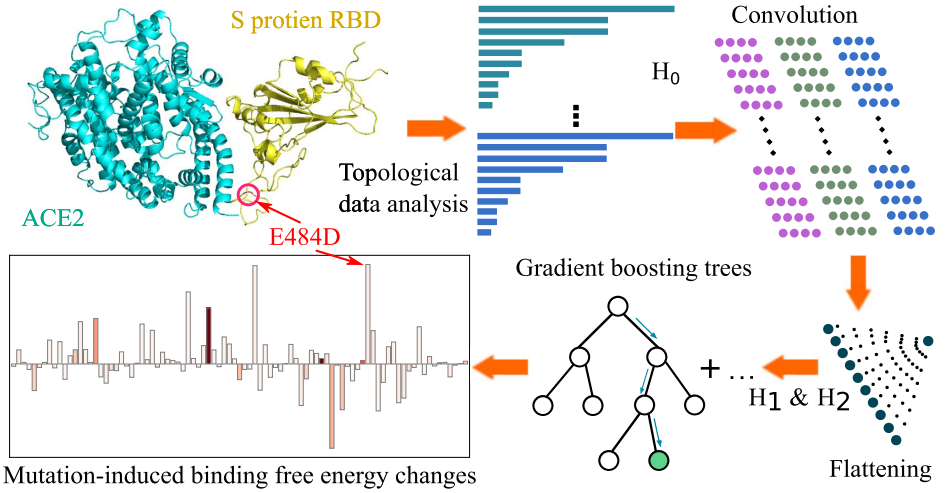

Figure 1: Illustration of the TopNetTree architecture for the prediction of mutation-induced ACE2 and S protein RBD binding free energy (BFE) changes [2]. H_k is the kth dimensional topological barcodes. Here, we will first take the number of H_0 as the input of the convolutional neural network (CNN). Next, the flatten layer of CNN, the number of H_1, H_2 will be used as the input features of the gradient boosting decision tree (GBDT) to predict the BFE changes from wild type to mutant type. The positive (negative) BFE changes strengthen (weaken) SARS-CoV-2 infectivity. A mutation at residue 484 from E (glutamic acid) to D (aspartic acid) near ACE2 is highlighted for its largest positive BFE change.

variants. Advanced mathematics offers some of the most powerful tools for such understanding [2, 13]. Although our method works better for both root mean square error (RMSE) and Pearson's r (R) compared to other methods [2], the limited size of training data related to SARS-CoV-2 may still affect the prediction of the BFE changes. However, we will still make efforts to collect more training data related to SARS-CoV-2, aiming to improve the performance of TopNetTree.

Acknowledgements

This work was supported in part by NSF Grants DMS-1721024, DMS-1761320, and IIS1900473, NIH grant GM126189, Bristol-Myers Squibb and Pfizer.

References

[1] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009. MR2476414

[2] J. Chen, R. Wang, M. Wang, and G.-W. Wei. Mutations strengthened SARS-CoV-2 infectivity. *Journal of Molecular Biology*, 432:5212–5226, 2020.

[3] J. Chen, R. Zhao, Y. Tong, and G.-W. Wei. Evolutionary de rham-hodge method. *Discrete and Continuous Dynamical Systems, Series B*, in press (doi: 10.3934/dcdsb.2020257), 2020.

[4] H. Edelsbrunner and J. Harer. Persistent homology-a survey. *Contemporary mathematics*, 453:257–282, 2008. MR2405684

[5] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, et al. SARS-CoV-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell*, 2020.

[6] B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*, 182:812–827.e19, 2020.

[7] Z. Meng, D. V. Anand, Y. Lu, J. Wu, and K. Xia. Weighted persistent homology for biomolecular data analysis. *Scientific reports*, 10(1):1–15, 2020.

[8] C. H. Rodrigues, Y. Myung, D. E. Pires, and D. B. Ascher. mcsm-ppi2: predicting the effects of mutations on protein–protein interactions. *Nucleic acids research*, 47(W1):W338–W344, 2019.

[9] R. Sanjuán and P. Domingo-Calap. Mechanisms of viral mutation. *Cellular and molecular life sciences*, 73(23):4433–4448, 2016.

[10] S. Sirin, J. R. Apgar, E. M. Bennett, and A. E. Keating. Ab-bind: Antibody binding mutational database for computational affinity predictions. *Protein Science*, 25(2):393–409, 2016.

[11] A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, and D. Veesler. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, 181:281–292.e6, 2020.

[12] M. Wang, Z. Cang, and G.-W. Wei. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence*, 2(2):116–123, 2020.

[13] R. Wang, Y. Hozumi, Y.-H. Zheng, C. Yin, and G.-W. Wei. Host immune response driving SARS-CoV-2 evolution. *Viruses*, 12:1095, 2020.

[14] R. Wang, D. D. Nguyen, and G.-W. Wei. Persistent spectral graph. *International Journal for Numerical Methods in Biomedical Engineering*, page e3376, 2020. MR4164275

[15] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, et al. A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269, 2020.

[16] K. L. Xia and G. W. Wei. Persistent homology analysis of protein structure, flexibility and folding. *International Journal for Numerical Methods in Biomedical Engineering*, 30:814–844, 2014. MR3247713

[17] C. Yin. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics*, 2020.

Jiahui Chen
Department of Mathematics
Michigan State University
MI 48824
USA
E-mail address: chenj159@msu.edu

Rui Wang
Department of Mathematics
Michigan State University
MI 48824
USA
E-mail address: wangru25@msu.edu

Guo-Wei Wei
Department of Mathematics
Michigan State University
MI 48824
USA
E-mail address: weig@msu.edu
URL: https://users.math.msu.edu/users/weig/

Received November 7, 2020