Vol. 12 (2026): Proceedings of the 2026 International Conference on Life Sciences and Health Engineering
(IC-LSHE 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/lshe.20260101

Predictive and Analytical Methods for Determining Viral Variation: Evolution and Frontier Breakthroughs

Mingshuo Zhang*

Department of Biotechnology, Yanbian University, Jilin, China

**Corresponding author: Mingshuo Zhang*

Abstract

The continuous mutation of viruses poses a persistent and severe challenge to global public health security, disease prevention and control, and biomedical research and development. The accurate prediction and in-depth analysis of viral mutation trends and their potential impacts are of critical strategic importance for constructing effective epidemic early warning systems and guiding the targeted development of vaccines and drugs. In recent years, the rapid development of artificial intelligence and high-throughput sequencing technologies has significantly enhanced our ability to monitor and infer viral evolutionary paths. In particular, the application of machine learning models in predicting mutation hotspots and functional impacts has opened new avenues for research. Moreover, interdisciplinary approaches combining genomics, structural biology, and computational modeling provide deeper insights into the mechanisms driving viral evolution. This paper aims to systematically review the evolution of methods for viral mutation prediction and analysis, delve into frontier technological breakthroughs in the field, objectively analyze the advantages and limitations of existing approaches, and provide an outlook on future directions. Furthermore, the integration of real-time surveillance data with predictive models is emphasized as a key factor in improving the timeliness and accuracy of mutation alerts. The goal is to offer a comprehensive and valuable reference for scientific research and practical applications in related fields, thereby assisting human society in better responding to the complex problems arising from viral variation.

Keywords

sequence alignment, phylogenetic tree, statistical model

1. Introduction

Since the global pandemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), although its severity and mortality rates have declined in some regions, the increasing emergence of new variants has led to frequent “breakthrough infections”-the phenomenon of reinfection in vaccinated or previously infected individuals. Rapid mutation of the virus not only presents significant obstacles to existing vaccines and therapeutic interventions but also markedly increases the potential risk of future pandemics. Consequently, the forward-looking prediction of viral mutations has become exceptionally important.

However, existing models for predicting SARS-CoV-2 mutations still face numerous challenges. A core difficulty lies in effectively integrating the regularity and inherent stochasticity of viral mutations while satisfying the real-world demands for minimal data and rapid response. From an evolutionary biology perspective, viral evolution is often characterized by “few-site mutations” and “rare beneficial mutations.” This implies that most mutations may be neutral or even deleterious, with only a very small fraction conferring advantages in aspects such as transmissibility or immune evasion, thereby prevailing through natural selection. The inability to accurately anticipate these critical evolutionary directions could result in greater harm to society.

To increase the accuracy of viral mutation prediction and promote the development of new models and methodologies, this paper consolidates multiple research findings to systematically summarize and review the current state of research. By synthesizing the experiences and lessons from existing methods, this paper hopes to provide clear guidance and a solid foundation for subsequent studies, ultimately contributing to the resolution of the global challenge of viral mutation. In recent years, data-driven approaches, particularly those combining large-scale genomic sequencing data with complex computational models such as machine learning, have demonstrated immense potential. Such methods can continuously monitor and analyze viral genomic sequences to capture early signals of potential high-risk variants, thereby securing valuable time for public health responses. In the future, the integration of multidimensional data and the development of more advanced artificial intelligence models will be crucial directions for the field.

2. Evolution of Traditional Methods

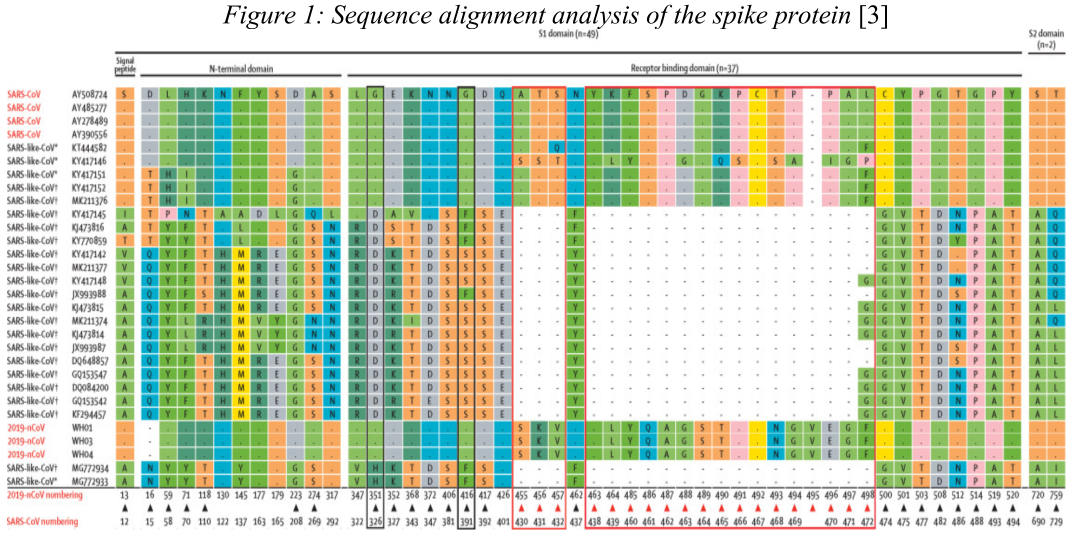
2.1 Analysis Methods Based on Sequence Alignment

One of the most widely used methods in viral analysis is the Basic Local Alignment Search Tool (BLAST). BLAST actually uses a local alignment algorithm that finds the matching segments with the highest similarity between two sequences. Its simple working flow includes the following steps.

It is used to discover the new coronavirus. In the whole process, researchers first acquire the complete genomic sequence of SARS-CoV-2 and then submit it as a query on the online website of the National Center for Biotechnology Information (NCBI) through the online BLAST tool. After that, the system will complete the following steps of algorithms to quickly identify similar sequences in the database. Finally, similar sequence alignment results can be obtained. The result will mark the similarity score, query coverage, E value (expected value), percent identity, etc., to represent the statistical similarity between them [1].

As shown in Figure 1, researchers can also use BLAST to perform comparative genomic analysis of coronavirus disease 2019 (COVID-19). Specifically, the ribonucleic acid genome of SARS-CoV-2 was compared with the RNA genomes of other human coronaviruses. Through the comparison of BLAST results, this study identified the highly conserved genomic regions of different coronaviruses and predicted the degree of viral immune escape [2].

Figure 1: Sequence alignment analysis of the spike protein [3]



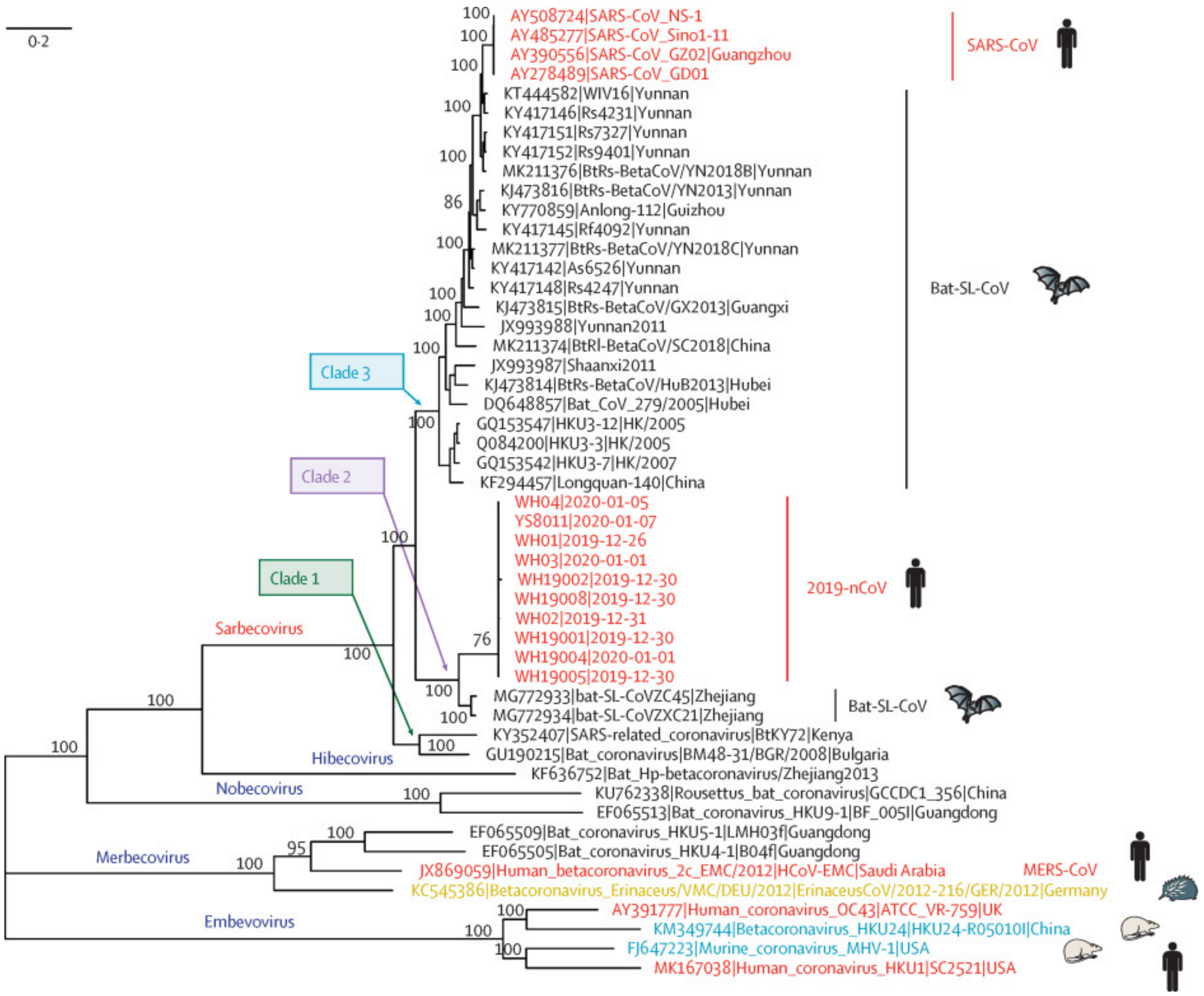
2.2 Phylogenetic Analysis

Phylogenetic analysis is a method used to identify the evolutionary path of a virus and predict its future behavior by visualizing the evolutionary relationships among different species (or other types of populations). A phylogenetic tree is a diagram showing the evolutionary relationships among various groups of organisms- be they species, subspecies, or even genes. The method principle is that the more similar the genes of the virus are, the closer they will be on the tree and the more likely they are to have diverged from a more recent common ancestor. There are generally three main steps involved in the construction of phylogenetic trees: sample collection and sequencing, multiple sequence alignment and tree inference.

This approach was applied in early studies of SARS-CoV-2 in Iran to determine the origin of this virus and the pathway of transmission, identification and typing of variant sites and inference of evolutionary relationships and divergence times. In this study, the authors constructed a phylogenetic tree on the basis of alignment of viral genomes from Iranian samples to a set of global reference genomes (e.g., all downloaded reference genomes from the GISAID database). Multiple nucleotide variant sites were subsequently obtained by aligning these sample sequences to a reference genome (e.g., Wuhan-Hu-1), as shown in Figure 2.

The topology of the obtained tree clearly shows that viral strains with specific mutations are clustered in one another and thus generate different evolutionary clades or lineages. Phylogenetic analysis can be applied to the temporal data of sample collection, and time-scaled Bayesian phylogenetic analysis can be used to reconstruct the divergence times of different lineages of viruses and the age of time to the most recent common ancestor (TMRCA) to understand the virus's evolutionary history and the transmission timeline of this virus in the human population [4].

Figure 2: Analysis of the phylogenetic tree of coronaviruses [3]


0·2
100
AY508724|SARS-CoV_NS-1
100
AY485277|SARS-CoV_Sino1-11
AY390556|SARS-CoV_GZ02|Guangzhou
100
AY278489|SARS-CoV_GD01
SARS-CoV
KT444582|WIV16|Yunnan
100
KY417146|Rs4231|Yunnan
KY417151|Rs7327|Yunnan
100
KY417152|Rs9401|Yunnan
MK211376|BtRs-BetaCoV/YN2018B|Yunnan
KJ473816|BtRs-BetaCoV/YN2013|Yunnan
86
KY770859|Anlong-112|Guizhou
KY417145|Rf4092|Yunnan
100
MK211377|BtRs-BetaCoV/YN2018C|Yunnan
KY417142|As6526|Yunnan
KY417148|Rs4247|Yunnan
100
KJ473815|BtRs-BetaCoV/GX2013|Guangxi
100
JX993988|Yunnan2011
100
MK211374|BtRl-BetaCoV/SC2018|China
Bat-SL-CoV
Clade 3
JX993987|Shaanxi2011
KJ473814|BtRs-BetaCoV/HuB2013|Hubei
DQ648857|Bat_CoV_279/2005|Hubei
100
GQ153547|HKU3-12|HK/2005
Q084200|HKU3-3|HK/2005
Clade 2
GQ153542|HKU3-7|HK/2007
KF294457|Longquan-140|China
WH04|2020-01-05
YS8011|2020-01-07
100
WH01|2019-12-26
WH03|2020-01-01
Clade 1
WH19002|2019-12-30
2019-nCoV
WH19008|2019-12-30
WH02|2019-12-31
Sarbecovirus
76
WH19001|2019-12-30
100
WH19004|2020-01-01
WH19005|2019-12-30
100
MG772933|bat-SL-CoVZC45|Zhejiang
100
100
MG772934|bat-SL-CoVZXC21|Zhejiang
Bat-SL-CoV
100
100
KY352407|SARS-related_coronavirus|BtKY72|Kenya
Hibecovirus
GU190215|Bat_coronavirus|BM48-31/BGR/2008|Bulgaria
KF636752|Bat_Hp-betacoronavirus/Zhejiang2013
Nobecovirus
100
KU762338|Rousettus_bat_coronavirus|GCCDC1_356|China
EF065513|Bat_coronavirus_HKU9-1|BF_005I|Guangdong
100
EF065509|Bat_coronavirus_HKU5-1|LMH03f|Guangdong
95
EF065505|Bat_coronavirus_HKU4-1|B04f|Guangdong
Merbecovirus
100
JX869059|Human_betacoronavirus_2c_EMC/2012|HCoV-EMC|Saudi Arabia
MERS-CoV
KC545386|Betacoronavirus_Erinaceus/VMC/DEU/2012|ErinaceusCoV/2012-216/GER/2012|Germany
100
AY391777|Human_coronavirus_OC43|ATCC_VR-759|UK
Embevovirus
100
KM349744|Betacoronavirus_HKU24|HKU24-R05010I|China
100
FJ647223|Murine_coronavirus_MHV-1|USA
MK167038|Human_coronavirus_HKU1|SC2521|USA

2.3 Statistical Models

The prediction and analysis of virus changes are not just hypothetical. They can be exact matches. Two methods are called maximum likelihood and Bayesian inference.

During the COVID-19 pandemic, scientists have studied virus family trees. This was very important. It helps track the virus in real time. It also helps predict dangerous new variants several weeks before normal health checks find them. By looking at the virus's family tree, researchers can see how the virus moves. They can also see how it changes from person to person. This gives real clues. These clues can help guide public health choices.

CoVerage is an automated tool. It was made by the Helmholtz Centre for Infection Research in Germany. It keeps watching how the SARS-CoV-2 virus has changed over time. It learns from how different versions of the virus spread among people. The system uses a math method named Bayesian inference. This method turns growth patterns into numbers. These numbers show how the virus spreads. In simple terms, raw gene data are used to obtain clear results. Public health workers can use these results directly [5].

3. Frontier Technology Breakthroughs

3.1 Multi-Omics Fusion Analysis

This method mixes different kinds of data. It uses data from genomics, transcriptomics, proteomics, and other similar fields. The goal is to learn how mutations truly work at a tiny level. For example, scientists can look at changes in the spike protein of a virus. They can also see how these changes affect RNA messages. Then, they run experiments to determine how tightly the virus binds. Together, this allows them to see more than just one type of data. However, this approach is not perfect. It's like trying to combine many different puzzles at once. Additionally, you need very powerful computers to make it work. In addition, sometimes, the results can surprise you.

3.2 Machine Learning and Deep Learning

Machine learning models, such as random forest and support vector machines (SVMs), have been used to forecast the antigenic changes of influenza viruses, but the performance of these models relies largely on feature engineering. On the other hand, deep neural networks (DNNs), long short-term memory (LSTM) networks, transformers and other deep learning approaches are capable of capturing complicated features automatically from raw data and have been applied to predict conformational changes in the SARS-CoV-2 S protein, for example. However, these models usually exhibit "black-box" characteristics and require large amounts of training datasets.

In a recent study entitled "Quantitatively characterizing the host adaptation process of SARS-CoV-2 variants through a deep learning-based prediction model" published in Cell Discovery, a team from the Chinese People's Liberation Army Center for Disease Control and Prevention, the Chinese Academy of Medical Sciences, Southeast University, and other institutions developed a model labeled ARNLE that can quantify the host adaptation process of SARS-CoV-2 variants and can accurately classify the viral host with high precision [6].

3.3 Molecular Dynamics (MD) Simulation

Molecular dynamics simulations are used to simulate the structural dynamics of viruses at the atomic level after mutation. The application is the study of mutations in the spike (S) protein of SARS-CoV-2 and how they affect the binding of antibodies (microlevel functional effects).

In 2022, scientists in Turkey employed computational methods such as molecular dynamics (MD) simulations to study 13 new SARS-CoV-2 mutations discovered in Turkey. Researchers have successfully characterized and studied the effects of newly discovered mutations on protein dynamics. This study offers an important theoretical basis for understanding the molecular mechanisms of viral variation and demonstrates the strong potential of MD simulations for rapidly assessing the potential effects of newly emerging mutations.

Effect: This method is mainly used to study the effects of new mutations on the structure, stability, and binding affinity of viral proteins to target host cell surface proteins such as angiotensin-converting enzyme 2 [7].

4. Future Directions

4.1 Fusion of Interdisciplinary Methods

This idea is about bringing together different areas, such as biology, computer science, mathematics, and physics. It uses the best parts of each method. For example, deep learning is good at finding features, and dynamic simulations can model objects at the molecular level. Combining these factors helps make better predictions for virus mutations. Mathematics can also help improve algorithms for integrating multiomics data. This makes data usage more efficient.

4.2 Real-time Dynamic Monitoring and Prediction Systems

Leveraging the power of rapid genetic sequencing, large-scale data systems, and machine learning, this study proposes a framework capable of tracking and forecasting viral mutations in real time. Such a system enables continuous collection and interpretation of global data on viral evolution, improving our capacity to recognize emerging patterns in how viruses change. Crucially, it supports public health decision-making by delivering timely, science-backed recommendations for outbreak response-for example, issuing early warnings when new variants emerge that demonstrate heightened transmissibility or immune evasion.

At the University of Nevada, Las Vegas, a research team developed an innovative tool named independent component analysis of variants, which integrates multifaceted data to improve the detection and surveillance of SARS-CoV-2 variants. Their approach involves regular wastewater sampling and analysis across diverse urban and rural settings. This offers a unique window into how the virus evolves within different communities. By applying independent component analysis (ICA), a computational method designed to separate blended signals, researchers can isolate and track specific mutation signatures-even within highly complex wastewater samples containing mixed variant populations. Between late 2021 and 2023, this system consistently identified major variants, including Delta, Omicron, and several recombinant forms [8].

4.3 Deep Fusion of Multiomics and AI

The continuous mutation of viruses poses a persistent and severe challenge to global public health security, disease prevention and control, and biomedical research and development. The accurate prediction and in-depth analysis of viral mutation trends and their potential impacts are of critical strategic importance for constructing effective epidemic early warning systems and guiding the targeted development of vaccines and drugs. This paper aims to systematically review the evolution of methods for viral mutation prediction and analysis, delve into frontier technological breakthroughs in the field, objectively analyze the advantages and limitations of existing approaches, and provide an outlook on future directions. The goal is to offer a comprehensive and valuable reference for scientific research and practical applications in related fields, thereby assisting human society in better responding to the complex problems arising from viral variation.

Since the global pandemic of SARS-CoV-2, although its severity and mortality rates have declined in some regions, the incessant emergence of new variants has led to frequent “breakthrough infections”-the phenomenon of reinfection in vaccinated or previously infected individuals. Rapid mutation of the virus not only presents significant obstacles to existing vaccines and therapeutic interventions but also markedly increases the potential risk of future pandemics. Consequently, the forward-looking prediction of viral mutations has become exceptionally important.

However, existing models for predicting SARS-CoV-2 mutations still face numerous challenges. A core difficulty lies in effectively integrating the regularity and inherent stochasticity of viral mutations while satisfying the real-world demands for minimal data and rapid response. From an evolutionary biology perspective, viral evolution is often characterized by “few-site mutations” and “rare beneficial mutations.” This implies that most mutations may be neutral or even deleterious, with only a very small fraction conferring advantages in aspects such as transmissibility or immune evasion, thereby prevailing through natural selection.

The inability to accurately anticipate these critical evolutionary directions could result in greater harm to society.

To increase the accuracy of viral mutation prediction and promote the development of new models and methodologies, this paper consolidates multiple research findings to systematically summarize and review the current state of research. By synthesizing the experiences and lessons from existing methods, this article hopes to provide clear guidance and a solid foundation for subsequent studies, ultimately contributing to the resolution of the global challenge of viral mutation. In recent years, data-driven approaches, particularly those combining large-scale genomic sequencing data with complex computational models such as machine learning, have demonstrated immense potential. Such methods can continuously monitor and analyze viral genomic sequences to capture early signals of potential high-risk variants, thereby securing valuable time for public health responses. In the future, the integration of multidimensional data and the development of more advanced artificial intelligence models will be crucial directions for the field.

5. Conclusion

This paper comprehensively surveys the evolution of predictive and analytical approaches for viral variation, ranging from conventional techniques such as sequence alignment, phylogenetic analysis, and statistical modeling to recent breakthroughs, including multiomics integration, machine and deep learning, and molecular dynamics simulations. This paper outlines the core principles, applications, strengths, and constraints of each method, offering a holistic perspective on the current state of viral mutation research.

The ongoing emergence of novel viral variants, such as SARS-CoV-2, highlights the urgent need for accurate and timely prediction of viral evolution. Although traditional approaches have established an essential foundation, the sheer complexity and speed of viral mutation demand more advanced tools. Emerging technologies-especially those combining large-scale genomic data with sophisticated computational models-provide unmatched potential for detecting subtle evolutionary patterns and anticipating the rise of high-risk variants.

Moving forward, the field is set to progress through the integration of interdisciplinary strategies, the creation of real-time dynamic prediction systems, and deeper merging of multiomics data with artificial intelligence. These synergistic approaches will refine our molecular-level insight into viral adaptation, increase the accuracy of forecasting models, and yield practical guidance for public health decision-making. Ultimately, enhancing our ability to predict and analyze viral changes will strengthen global health security, speed up the development of precise vaccines and antivirals, and improve our capacity to curb the effects of future pandemics. Such scientific advancements are indispensable for humanity’s proactive stance against the enduring threat of viral evolution.

References

[1] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. BLAST+: architecture and applications. BMC Bioinformatics. 2009, 10(1), p. 421. https://doi.org/10.1186/1471-2105-10-421.

[2] Hussein, M., Andrade dos Ramos, Z., Berkhout, B. and Herrera-Carrillo, E. In Silico Prediction and Selection of Target Sequences in the SARS-CoV-2 RNA Genome for an Antiviral Attack. Viruses. 2022, 14(2), p. 385. https://doi.org/10.3390/v14020385.

[3] Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. 2020, 395(10224), pp. 565-574. https://doi.org/10.1016/S0140-6736(20)30251-8.

[4] d’Alessandro, M., Bergantini, L., Cameli, P., Curatola, G., Remediani, L., Sestini, P. and Bargagli, E. Peripheral biomarkers’ panel for severe COVID-19 patients. Journal of Medical Virology. 2021, 93(3), pp. 1230-1232. https://doi.org/10.1002/jmv.26577.

[5] Norwood, K., Deng, Z.-L., Reimering, S., Robertson, G., Foroughmand-Araabi, M.-H., Goliaei, S., Hölzer, M., Klawonn, F. and McHardy, A. C. In silico genomic surveillance by CoVerage predicts and characterizes SARS-CoV-2 variants of interest. Nature Communications. 2025, 16(1), p. 6281. https://doi.org/10.1038/s41467-025-60231-4.

[6] Li, C., Chen, L. and Lan, T. Artificial intelligence (AI) reveals the pandemic potential and host adaptation of SARS-CoV-2 variants. Cell. 2024, 2(15), pp. 1152–1166.

[7] Unlu, S., Uskudar-Guclu, A. and Cela, I. The impacts of 13 novel mutations of SARS-CoV-2 on protein dynamics: In silico analysis from Turkey. Human Gene. 2022, 33, p. 201040. https://doi.org/https://doi.org/10.1016/j.humgen.2022.201040.

[8] Zhuang, X., Vo, V., Moshi, M. A., Dhede, K., Ghani, N., Akbar, S., Chang, C.-L., Young, A. K., Buttery, E., Bendik, W., et al. Early detection of emerging SARS-CoV-2 Variants from wastewater through genome sequencing and machine learning. Nature Communications. 2025, 16(1), p. 6272. https://doi.org/10.1038/s41467-025-61280-5.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access