Article

Not peer-reviewed version

Mutation Sites Increase over Time in SARS-CoV-2 Variants

Liaofu Luo * and Jun Lv *

Posted Date: 14 November 2024

doi: 10.20944/preprints202411.0947.v1

Keywords: SARS-CoV-2; variant; macro-lineage; evolution timeline

 doi:10.20944/preprints202411.0947.v1

Article

Mutation Sites Increase Over Time in SARS-CoV-2 Variants

Liaofu Luo [1,*] and Jun Lv [2,*]

[1] Faculty of Physical Science and Technology, Inner Mongolia University, 235 West College Road, Hohhot 010021, China; lolfcm@imu.edu.cn

[2] College of Science, Inner Mongolia University of Technology, 49 Aymin Street, Hohhot 010051, China; lujun@imut.edu.cn

* Correspondence: lolfcm@imu.edu.cn (L.L.); lujun@imut.edu.cn (J.L.); Tel.: +86-471-4992676 (L.L.); +86-13039500654 (J.L.)

Abstract: The evolution timeline of SARS-CoV-2 is examined. We found an approximately linear relationship between the number of mutated sites (x) on the spike protein of a variant and its first global sample collection time. By combining the emergence of novel strains at a given x with this linear relationship, we can predict the emergence time of macro-lineages. It is forecasted that macro-lineage Q will emerge shortly after the emergence of lineage P.

Keywords: SARS-CoV-2; variant; macro-lineage; evolution timeline

1. Introduction

The continuous spread of the novel coronavirus over the past four years has been driven by successive waves of SARS-CoV-2 mutations. Predicting the generation of new strains and the emergence of multiple macro-lineages remains a significant challenge. A network-based inference approach has been proposed for short- to mid-term predictions [1]. Given the crucial role of spike protein mutations in the rapid evolution of SARS-CoV-2, deep learning methodologies have been suggested to predict future protein sequences, leveraging Large Language Models [2–5]. These models have shown promise in training protein language models for forecasting pandemic-related protein mutations. However, ensuring the continuous incorporation of new experimental data and accurately predicting the future trajectory of macro-lineages remains a critical challenge.

We previously introduced a mathematical model to analyze the dynamics of COVID-19 spread [6]. Using a reconstructed phylogenetic tree and the A-X model, a statistical approach for generating new strains, we combined a set of existing mutation sites (A) with a set of randomly generated sites (X) to model the emergence of new strains on the phylogenetic tree and explain the patterns of multiple SARS-CoV-2 macro-lineages [7,8]. By expanding the stochastic sampling to a larger scale, we uncovered the statistical principles governing the emergence of new strains. Our findings show that the probability of a macro-lineage's emergence is related to the number x of randomly generated sites within the X set. As x increases, the proportions of macro-lineages change: lineage O surpasses lineage N, followed by lineage P surpassing lineage O, and ultimately, lineage Q surpassing lineage P. We initially predicted the emergence of macro-lineage P, which has since been observed. Furthermore, we forecasted the emergence of macro-lineage Q when x reaches a sufficiently large value. These results provide a crucial theoretical framework for understanding the evolution of SARS-CoV-2.

However, the precise timeline of SARS-CoV-2 evolution remains unknown. To predict the future trajectory of macro-lineage P and the emergence of macro-lineage Q, it is essential to understand how the number of mutated sites (NMS) for selected SARS-CoV-2 variants and the accumulated number of mutated sites (ANMS) on the spike protein evolve over time. This is the primary motivation for the present study. Despite the complexity of the various factors influencing viral evolution, we identified an approximate linear relationship between the number of mutated sites for a given variant

 doi:10.20944/preprints202411.0947.v1

(i.e., x) and its worldwide first sample collection date (i.e., physical time t). Consequently, this enables us to predict the timeline of macro-lineage transformations. To further enhance our understanding of SARS-CoV-2 evolution, the conditions for establishing this linear relationship and its connection to the A-X model are also discussed in this manuscript.

2. Materials and Methods

2.1. Materials

The SARS-CoV-2 mutants are listed in Table 1 in chronological order of their worldwide first sample collection dates. The number of mutated sites and the total number of accumulated mutations on the spike protein are also provided. Characteristic mutations for a lineage are defined as nonsynonymous substitutions or deletions occurring in more than 75% of sequences within that lineage. The data are sourced from outbreak.info [9].

Table 1. SARS-CoV-2 variants (numbers of mutated sites on spike protein).

Macro-lineage	Variant	NMS *	ANMS †	Earliest date ‡	Variant	NMS *	ANMS †	Earliest date ‡
N-lineage	B.1	1	1	15 Jan 2020	B.1.621	9	40	19 Sep 2020
	B.1.177	2	2	7 Mar 2020	C.37	14	52	8 Nov 2020
	P.2	3	4	15 Apr 2020	B.1.526	4	53	15 Nov 2020
	B.1.1.7	10	13	14 May 2020	B.1.525	9	57	11 Dec 2020
	B.1.429	4	16	6 Jul 2020	P.3	7	59	15 Jan 2021
	B.1.351	10	23	9 Jul 2020	AZ.2	6	60	5 Feb 2021
	B.1.617.2	9	29	7 Sep 2020	AV.1	10	64	23 Mar 2021
	P.1	12	36	11 Sep 2020	B.1.1.529	7	67	15 Apr 2021
	B.1.617.1	5	37	15 Sep 2020	C.1.2	15	71	11 May 2021
O-lineage	BA.1	33	87	27 Jan 2021	BN.1.2	40	106	7 Feb 2022
	BA.1.1	35	88	28 Jan 2021	CH.1.1	41	106	12 May 2022
	BA.2	31	96	25 Mar 2021	XBB.1.5	42	111	12 Jun 2022
	BA.2.12.1	33	97	28 Sep 2021	BM.4.1.1	39	111	20 Jul 2022
	BA.2.65	31	97	11 Oct 2021	CH.1.1.1	42	112	15 Oct 2022
	BA.1.1.15	37	97	27 Nov 2021	XBB.1.16	43	113	4 Jan 2023
	BA.5	34	98	9 Dec 2021	EG.1	43	114	16 Jan 2023
	BA.4.1	35	99	14 Dec 2021	HV.1	46	115	29 Jan 2023
	BQ.1.1	37	101	20 Dec 2021	HK.3	45	116	29 Jan 2023
	BA.2.75	30	103	31 Dec 2021	EG.5.1	44	116	31 Jan 2023
	BF.5	35	104	8 Jan 2022	DV.7.1	45	117	29 May 2023
	BF.7	35	104	24 Jan 2022				
P-lineage	JN.1	60	132	13 Jan 2023	XDQ.1	55	139	5 Jan 2024
	BA.2.86.1	59	132	17 Jan 2023	KP.3	63	139	7 Jan 2024
	BA.2.86	58	132	11 Mar 2023	LB.1	64	139	15 Jan 2024
	JN.2	59	132	22 Jun 2023	KP.1	63	140	1 Feb 2024
	JN.1.7	62	134	25 Sep 2023	KS.1	58	142	15 Feb 2024
	JN.1.11.1	62	135	29 Dec 2023	KP.1.1.3	65	142	23 Feb 2024
	KP.3.1.1	64	136	1 Jan 2024	XDV.1	56	142	26 Feb 2024
	KP.2	59	136	2 Jan 2024	LP.1	66	143	22 Apr 2024
	JN.1.37	61	137	3 Jan 2024	XED	64	144	19 Jun 2024
	XEB	61	138	3 Jan 2024	XEC	65	145	28 Jun 2024

* NMS: number of mutated sites. † ANMS: accumulated number of mutated sites. ‡ Dates are based on the worldwide first sample collection date.

2.2. Methods

The least squares regression analysis is conducted to examine the relationship between the number of mutated sites and the first sample collection date of the variants. In general, for a

 doi:10.20944/preprints202411.0947.v1

dependent variable y and an independent variable x, with observed values y_i at $x=x_i$, the linear regression equation is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \quad (1)$$

where $\hat{y}$ is the predicted value from the model, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the regression coefficients. The standard error of the prediction $SE(\hat{y})$ is calculated as

$$SE(\hat{y}) = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \quad (2)$$

where $\bar{x}$ is the mean of the x_i values, and n is the number of samples. The standard error of the slope $\hat{\beta}_1$ is

$$SE(\hat{\beta}_1) = s\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}. \quad (3)$$

The standard error of the intercept $\hat{\beta}_0$ is

$$SE(\hat{\beta}_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}. \quad (4)$$

In equations (2)-(4), the term s refers to the Residual Standard Error (RSE), also known as the model's sigma, which is defined as

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}}. \quad (5)$$

The 95% confidence interval is also used to define a range of parameter estimates that includes the true value with 95% probability. The confidence interval is calculated as

$$CI(\hat{q}) = \hat{q} \pm t(\alpha/2, df)SE(\hat{q}), \quad (6)$$

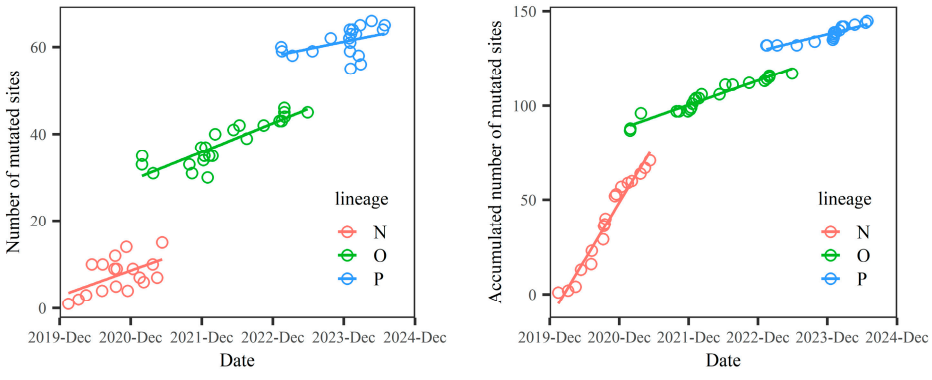
where $\hat{q} \in \{\hat{y}, \hat{\beta}_1, \hat{\beta}_0\}$, and $t(\alpha/2,df)$ is the critical value from the t-distribution with a confidence level of α and degrees of freedom df. For a 95% confidence interval, α=5%.

The method is assessed using the R-squared (R^2) and the Residual Standard Error (RSE). The model's goodness of fit is indicated by a high R^2 (close to 1) and a low RSE.

3. Results

3.1. Each Macro-Lineage Has a Specific Survival Time. The Relationship Between the Number of Mutated Sites and Time t Is a Discontinuous Function

Let the number of mutated sites on the spike protein (NMS) for selected SARS-CoV-2 variants be denoted as x, where x is a function of time, $x=x(t)$. The accumulated number of mutated sites on the spike protein (ANMS) at time t is represented as $s(t)$. We found that both $x(t)$ and $s(t)$ are discontinuous functions of time, corresponding to three macro-lineages, as shown in Figure 1. Figure 1 is based on the data presented in Table 1.



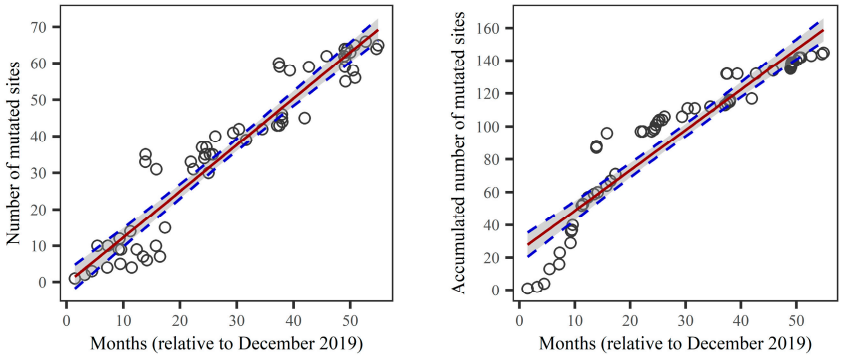
(**a**) (**b**)

 doi:10.20944/preprints202411.0947.v1

Figure 1. Number of mutated sites x (**a**) and accumulated mutated sites s (**b**) in the spike protein as a function of the first sample collection date t for the variant.

3.2. Linear Regression of the Number of Mutated Sites in a Variant Versus Sample Collection Date as a Good Approximation

The linear regression of the relationship between the number of mutated sites and the first sample collection date was performed, and the results are shown in Figure 2. The left panel (Figure 2a) depicts the evolution of the number of mutated sites (NMS), while the right panel (Figure 2b) shows the evolution of the accumulated number of mutated sites (ANMS). We found that the linear regression of the number of mutated sites in the variant provides a good approximation of the increasing trend in the number of mutated sites during viral evolution. The R-squared (R^2) and the Residual Standard Error (RSE) for the linear regression of NMS are R^2=0.91 and RSE=6.51, respectively. Furthermore, using Equations (3) and (6), we obtain a slope of the regression line dx/dt=1.268 per month, with a 95% confidence interval of ±0.103. For the linear regression of ANMS, the estimate yields R^2=0.88 and RSE=14.66. The slope of the regression line for ANMS is ds/dt=2.452 per month, with a 95% confidence interval of ±0.232. Therefore, the linear regression for NMS provides a better fit than for ANMS, and we will use the linear relationship between the number of mutated sites x and collection time t to predict the emergence of new lineages.



(**a**) (**b**)

Figure 2. Linear regression of the number of mutated sites (**a**) and the accumulated number of mutated sites (**b**) versus the first sample collection date. The 95% confidence intervals, calculated using Equations (2) and (6), are represented by the blue dotted lines.

3.3. Prediction of the Emergence of the Q Macro-Lineage

Building on the studies above, we are able to predict the emergence of the new macro-lineage Q. The emergence of Q was initially forecasted in references [7,8], where the A-X model was proposed to generate new strains on the phylogenetic tree. The core concept of the model involves combining set A (existing mutated sites) with set X, which contains x randomly generated sites, to predict how a novel strain is generated on the tree. By expanding stochastic sampling to a larger scale, statistical laws governing new strain production and the probability of macro-lineage (PML) versus x can be derived. These analyses were based on data from 36 to 40 mutants. To improve statistical accuracy, we now apply the enlarged dataset of 61 mutants (Table 1) and the same stochastic sampling method as described in reference [8]. The results of PML versus x are presented in Figure 3.

From Figure 3, the demarcation values (99% percentile) are as follows:

New-mutant ∈ N when $x \le 20$; New-mutant ∈ N or O when $21 \le x \le 30$; New-mutant ∈ O when $31 \le x \le 37$; New-mutant ∈ O or P when $38 \le x \le 62$; New-mutant ∈ P when $63 \le x \le 78$; New-mutant ∈ P or Q when $79 \le x \le 107$; New-mutant ∈ Q when $x \ge 108$.

Based on Equations (1) (2) and (6), we predict that the number of mutated sites will reach x=78.25±3.76 to 79.52±3.85 at the 62nd to 63rd month, and x=108.69±6.07 at the 86th month, starting from December 2019. Combining these data with the demarcation values x_1=79 (for Q's initial emergence)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 November 2024 doi:10.20944/preprints202411.0947.v1

and x_2=108 (for a strong outbreak of Q) from Figure 3, we forecast that the macro-lineage Q will emerge around February 2025 (x~79) and, after approximately 23 months, will reach the stage of a strong outbreak (x~108).

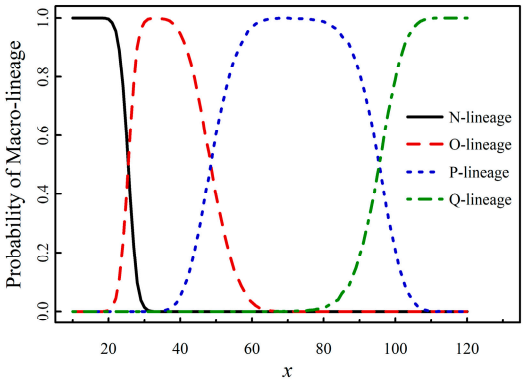


Figure 3. PML(Probability of Macro-lineage) versus x for 61 mutants. PMLs are calculated by use of sequence data in Table 1.

4. Discussion

4.1. Why Is the Increase in the Number of Mutated Sites of a Variant Approximately Linear with Respect to Its Emergence Time?

Figure 1 illustrates that the relationship between the number of mutated sites (NMS) and time t is discontinuous, with a stepwise change occurring at the point of lineage transformation. However, the slope of the NMS increase between two neighboring lineages also varies. Generally, the new lineage exhibits a lower slope of increase compared to the older lineage, which compensates for the stepwise change at the lineage transformation. This explains why the relationship between NMS and the emergence time of a variant is approximately linear. Our prediction for the Q lineage is based on this linear relationship, which can be extended to longer time periods.

4.2. What Is the Relationship Between the Time Prediction for the Emergence of a Macro-Lineage and the Mutant Prediction in the A-X Model?

In this article, we forecast the timeline for mutant evolution, while the A-X model primarily focuses on the stochastic generation of mutants on the phylogenetic tree. The time required for the emergence of a mutant and the number of randomly generated sites within a mutant are intrinsically linked. Therefore, in section 3.3, the prediction for the Q lineage is derived using the A-X model. In fact, the emergence of any new lineage can be independently predicted, provided there is sufficient data on the survival time of macro-lineages.

5. Conclusions

This manuscript extends the work presented in the articles "An Evolutionary Theory on Virus Mutation in COVID-19" [7] and "Prediction on the Emergence of SARS-CoV-2 Based on Evolutionary Theory of Virus Mutation" [8]. The main arguments and conclusions are summarized as follows:

1. The n-distance algorithm, applied in UPGMA, generates a phylogenetic tree of viral evolution based on amino acid mutations in the spike protein. The reconstructed tree aligns closely with established evolutionary data;
2. The A-X model is introduced to simulate the generation of new strains on the phylogenetic tree. By combining set A (existing mutated sites) with set X (which includes x randomly generated sites), we can predict the emergence of novel strains. Expanding stochastic sampling to a larger scale reveals statistical patterns governing new strain production. As x increases, the proportions of the four macro-lineages change: lineage O surpasses N first, followed by lineage P surpassing O, and finally, lineage Q emerges;

 doi:10.20944/preprints202411.0947.v1

3. A linear regression between the number of mutated sites (NMS) for a variant (i.e., x) and its worldwide first sample collection time (i.e., t) provides a good approximation. This linearity arises from the combined effects of stepwise changes in NMS at lineage transformations and varying slopes of NMS versus time in neighboring lineages;
4. By integrating the information on novel strain production at a given x from the A-X model and the linear relationship between x and t, we forecast that macro-lineage Q will emerge around February 2025 (when x≈79), and will reach a stage of strong outbreak approximately 23 months later (when x≈108).

Author Contributions: Conceptualization, L.L.; validation, J.L., and L.L.; investigation, J.L.; writing—original draft preparation, L.L.; writing—review and editing, J.L.; visualization, J.L.; supervision, L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgements: We sincerely appreciate Dr. Ying Zhang for her assistance with data collection, insightful discussions, and valuable suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rahnsch, B.; Taghizadeh, L. Network-based uncertainty quantification for mathematical models in epidemiology. *J. Theor. Biol.* **2024**, *577*, 111671. https://doi.org/10.1016/j.jtbi.2023.111671.
2. Ramachandran, A.; Lumetta, S.S.; Chen, D. PandoGen: generating complete instances of future SARS-CoV-2 sequences using deep learning. *PLoS Comput. Biol.* **2024**, *20*, e1011790. https://doi.org/10.1371/journal.pcbi.1011790.
3. Fowler, D.; Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **2014**, *11*, 801–807. https://doi.org/10.1038/nmeth.3027.
4. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381.
5. Ferruz, N.; Schmidt, S.; Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **2022**, *13*, 4348. https://doi.org/10.1038/s41467-022-32007-7.
6. Luo, L.; Lv, J. Mathematical modelling of virus spreading in COVID-19. *Viruses* **2023**, *15*, 1788. https://doi.org/10.3390/v15091788.
7. Luo, L.; Lv, J. An evolutionary theory on virus mutation in COVID-19. *Virus Res.* **2024**, *344*, 199358. https://doi.org/10.1016/j.virusres.2024.199358.
8. Luo, L.; Lv, J. Prediction on emergence of SARS-CoV-2 based on evolutionary theory of virus mutation. Available online: https://ssrn.com/abstract=4938698 (accessed on 31 August 2024).
9. Gangavarapu, K.; Latif, A.A.; Mullen, J.L.; Alkuzweny, M.; Hufbauer, E.; Tsueng, G.; Haag, E.; Zeller, M.; Aceves, C.M.; Zaiets, K.; et al. Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nat Methods* **2023**, *20*, 512–522. https://doi.org/10.1038/s41592-023-01769-3.