Evolutionary trajectory of SARS-CoV-2 genome

Anyou Wang (✉ anyou.wang@alumni.ucr.edu)

University of Memphis <https://orcid.org/0000-0002-4981-3606>

Biological Sciences - Article

Keywords:

Posted Date: October 25th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1009010/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Evolutionary trajectory of SARS-CoV-2 genome

Anyou Wang ORCID 0000-0002-4981-3606

Feinstone Center for Genomic Research, University of Memphis, Memphis, TN 38152, USA

Contact:

anyou.wang@alumni.ucr.edu

Project website

<https://combai.org/ai/covidgenome/>

Abstract: Traditionally alignment-based phylogenetics faces challenges to uncover the evolutionary trajectory of SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2). This study develops a novel alignment-free system and reveals the evolutionary trajectory of SARS-CoV-2 from more than one million of genome sequences. This new system contains Fréchet distance(Fr) and artificial recurrent neural network. Fr computes the distance between variant and reference genome, which is decomposed into 84 features (4 single nucleotides, 16 dinucleotides and 64 codons). Recurrent neural network predicts and forecasts time-series Fr trajectory, inferring SARS-CoV-2 evolutionary trajectory. Generally SARS-CoV-2 genome mutates rapidly via deletion during COVID-19 pandemic. Among single nucleotides, C mutates fast but T changes slowly. C-prefix dinucleotide (e.g. CG and CT) also loses dramatically during evolution. Similarly, the virus genome also deletes several codons prefixed by C (e.g. CCT) but gains several T and A prefix codons (e.g.TTA and ATT) during its evolution. Interestingly, codon CCT and CT centrally control the entire SARS-CoV-2 genome, and their evolutionary trajectories fit COVID-19 cases spike. Therefore C-prefix feature trajectory marks SARS-CoV-2 evolution.

Key word:

SARS-CoV-2; COVID-19; Fréchet distance; artificial recurrent neural network; genome; evolutionary trajectory

Introduction

SARS-CoV-2 causes global COVID-19 pandemic¹⁻⁶. Understanding its evolutionary trajectory helps to combat this pandemic.

Phylogenetics tree has been widely applied to evolution studies and has generated rich useful information on virus evolution^{3,7-9}. However, when millions of SARS-CoV-2 genome sequences merges¹⁰, this conventionally theory and approach sounds folded. More importantly, this traditional method depends on alignment⁹, which aligns a similar portion and ignores major different parts. For example, the traditional approach aligns two sequences p and q, in which p poses 8A (AAAAAAAA) and q has 10A (AAAAAAAAAA). The result comes out as 100% similarity and it automatically ignores another two AA in q. Actually, these two AA contributes variations of these two sequences like codon capacity(3 AAA codons in q but only 2 AAA codons in p). Therefore, the conventional phylogenetics actually finds a limited number of minor variations inside the aligned-similar sequences(e.g. a single nucleotide mutation in the highest similar fragment), and it disregards the unaligned section. This unaligned proportion may contain major variations of sequences and pose primary biological functions. Therefore alignment-based methods certainly underestimate a certain number of biological meaningful patterns.

Alignment-free methods have been developed¹¹, but they have focused on either nucleotide positions or contents and they do not have enough features to describe the genome. More importantly, these

approaches lack robustness to buffer noises from sequencing experiments. These current alignment-free approaches have few applications.

A practical alignment-free approach should be sensitive enough to detect variations of sequences and be robust enough to buffer noises from sequence variations. This present study developed a novel sensitive and robust system, which includes Fréchet distance (Fr)¹² and long short-term memory¹³(LSTM, a recurrent neural network architecture). Fr can be implemented to count on not only positions of a feature but also its position order as well as its contents. Fr is sensitive to detect sequence variations, but noises still remain. Here LSTM plays its role. LSTM catches the key useful information from noised data and creates robust patterns. This Fr-LSTM system has successfully generated evolutionary trajectories of SARS-CoV-2 genome in this study and evolutionary trajectory and origin of SARS-CoV-2 variant in another parallel study¹⁴.

Results

Method development for investigating virus genome evolutionary trajectory

To understand virus genome evolutionary trajectory, this study decomposes the 30k SARS-CoV-2 genome into 84 genome features, including 4 single nucleotides, 16 dinucleotides and 64 codons, and then examines the evolutionary trajectories of these 84 features to infer the virus genome evolutionary trajectory(materials and methods). These 84 feature dynamic Frs were computed from a total 1,128,954 genome sequences filtered out from 2,212,864 samples available in GISAID (<https://www.gisaid.org/>) on July 4, 2021, when the sequence data was downloaded (materials and methods).

To measure the evolutionary trajectory of a given genome feature, this study computed its distance from the corresponding feature in SARS-CoV-2 reference genome(NC_045512.2)⁶ at one time point and then appended these distance together following time order, constructing a dynamic distance profile.

At a given time point, this study treats the positions of a given feature in a reference sequence as a trajectory. For example, the positions of a single nucleotide A in a reference sequence of AAAAAAAAAA can be recorded as [0,1,2,3,4,5,6,7], which constructs a reference trajectory. Besides positions, A's content in this curve is 100%. Similarly, the positions of a given feature on a variant genome also construct a mutant trajectory. Computing the distance of these two trajectories can infer the evolutionary distance of this feature.

This study employs the coupling discrete Fr¹² to measure the distance of two trajectories. Fr takes care of ordered positions and its implementation can embrace the contents of these positions (materials and methods). For example, the Fr for a single nucleotide A between two sequences(reference: AAAAAAAAAA with ordered positions[0,1,2,3,4,5,6,7], variant: AAAAAAAAAA with ordered positions[0,1,2,3,4,5,6,7,8,9]) is 2 because its content in two sequences is the same (100% for both) and its Fr becomes the maximum distance of its ordered positions. This distance is equal to the absolute value of difference between the position #10 in the variant and the position #8 in the reference, which is 2. Similarly, the Fr for dinucleotide and codon can be computed in the same way (Figure 1A).

Fr is always positive, but when the number of positions for a given feature in a variant is less than that of reference, this study adds a negative sign to Fr to mark it as deletion(loss). Automatically, a feature with positive distance is regarded as insertion(gain).

To get the first impression of SARS-CoV-2 Fr distribution, this study compared the 84 feature Frs of three groups, alpha, global and China. Alpha variant (EPI_ISL_601443) sequence was used to compute Frs for alpha group, and global and China group denoted the Fr median of all global samples and Chinese samples in 2019 respectively(Figure 1B). As expected, most features of the China group had Fr closed to 0. These sample genomes were very similar to reference. In contrast, most feature Frs for alpha variant and global variants moved far away from 0, indicating these features have dramatically mutated. In addition, most feature Frs distributed below 0, indicating that deletion dominates SARS-CoV-2 mutations.

To understand the overall evolutionary trajectory of SARS-CoV-2 genome, this study examined the daily dynamic trend of the Fr median for all features across all samples from the UK(Figure 1C) because UK was the country who had completely collected virus sequences from virus emerge. Virus sequences have been generated from varied conditions and sequence noises were unavoidable. To diminish the noise effects and appreciate a clear trend of evolutionary trajectory from December 2019 to July 4, 2021 and to forecast its trends beyond July 4, 2021, this study employed LSTM. A LSTM model with 4 layers and two dropouts was built (materials and methods) and this model was applied to all predictions and forecasts in this study. The start point to train this model was set to 21 days (input-width=21) and the forecast days was set to 30 days after July 4, 2021. The test size was set to 5% of days. The walk forward strategy of one position shift was applied when modeling. The median Fr as a single variable was used to train and predict Fr value in this study unless specific notice, and the predicted value was used as the metric to explain all results in this study.

The Fr median for all SARS-CoV-2 genomes declined below 0 (Figure 1C), indicating its genome generally shortens with time. From 3/2020 to 9/2020, the virus genome completed its first trajectory of

gradual deletion, with Fr declined gradually from -150 to -230, but it suffered the biggest loss from 12/2020 to 5/2021, with Fr dropping from -225 to -320. SARS-CoV-2 continues to shorten its genome in the near future as forecasted after 7/2021. Therefore, SARS-CoV-2 overall evolves to shorten its genome.

To predict the infection cases of SARS-CoV-2, this study used the same LSTM model described above but multivariate (84 features) matrix was used as training matrix and the number of infection cases was used as the response. The predicted cases fitted the actual case very well (Figure 1D) and the infection cases were corresponding to the virus genome evolutionary trajectory (Compared Figure 1C vs Figure 1D), indicating that virus genome composition can predict the infection cases. Interestingly, the predicted cases after May, 2021 was much higher than actual cases, suggesting that vaccination may knock down the natural cases and genome sequences after 5/2021 may be biased to unvaccinated patients.

Together, SARS-CoV-2 shortens its genome to infect human objects. The data above confirmed that the system established here works for virus genome evolution study.

Evolutionary trajectory of single nucleotide

To appreciate the evolutionary trajectory of single nucleotides, this study investigated the dynamic changes of Fr median of all global samples(**Figure 2**). Among 4 single nucleotides, C mutated easily, and its Fr dropped from -40 to -120 between 1/2020 and 5/2020 and then gradually declined until another sharp drop (max Fr = -168) from 5/2021(**Figure 2A**). These two declined trajectories were respectively corresponding to the global pandemic in May, 2020 and delta variant spread in May 2021,

indicating that C plays a key role in SARS-CoV-2 evolution. Following C mutation degree, A was the next (max Fr = -137) and then G (Max Fr = -107). T was the most stable nucleotide (max Fr = -104) in SARS-CoV-2 genome(**Figure 2B-D**).

Dinucleotide evolutionary trajectory

Among dinucleotide evolution, the biggest deletion occurred in CG(**Figure 3A**). CG stayed at stable state as wild type before 3/2020, but it had a sharp deletion trajectory beginning 3/2020 to 5/2020 and then gradually shortened to max (Fr = -880) in 3/2021(**Figure 3A**). This CG trajectory was corresponding to the first global COVID-19 pandemic curve. GC and CT also went through deletion, but their trajectories were different(**Figure 3B-C**). GC posed two Fr drops in 7/2020 and 3/2021 but the whole trajectory was rich with up and down noises around the prediction curve. However, CT began its biggest deletion from 01/2020(Fr from -60 to -160) and then continued another long deletion trajectory from 12/2021 to the near future as forecasted. The whole CT trajectory was much clearer than GC, and this CT trajectory was also corresponding to the wave of virus global outbreak.

In contrast to the continued deletion trajectories of most dinucleotides, AT stayed very stable during pandemic until 05/2021, when it also went through a slight deletion during delta variant outbreak(**Figure 3D**), suggesting that delta variant even destroys the its own stable system for widening infection. Overall, all dinucleotides undergo deletions during evolution(**Figure 3E**, complete plots was shown in this project website<https://combai.org/ai/covidgenome/>), but C-prefix dinucleotides including CG, CC, CA and CT experience the biggest deletion while the AT and TG stay stable during pandemic.

Codon evolutionary trajectory

In contrast to the loss trajectory of all dinucleotides, trinucleotide codons hosted several members of gain during evolution(Figure 4 A-C). Three typical gain trajectories were observed, including gradual increase(e.g. TTA, Figure 4A), stable gain(e.g. GGG, Figure 4B), and waving gain(ATT, Figure 4C). Similarly, SARS-CoV-2 genome also experienced gradual loss(e.g. GCT, Figure 4D), waving deletion (e.g. CCT, Figure 4E), stable loss (e.g. CGT, Figure 4F), and chaotic loss (e.g. Figure 4G).

Among these loss and gain trajectories, CCT had three clear deletion periods(1/2020, 9/2020, and 6/2021) and its trajectory was very clear and stable during pandemic stable periods but was waving up and down around prediction curve, fitting the three waves of virus global outbreaks. This indicated its important role in the pandemic. Generally, several codons prefixed with C such as CGA, CCC, CGG, CCA suffered big loss during evolution(Figure 4 H).

The most import features in SARS-CoV-2 genome

To understand the crucial features in the SARS-CoV-2 genome, this study employed FINET¹⁵ to build an interaction network of these 84 feature Frs by using all global samples and searched the most important hubs in this network. Three C-prefix features (CCT, CT and CAA) worked as the most important player in this network(**Figure 5**). Especially, CCT strongly controlled CC that further mediated GCC, AND, AAC and ACC. Recalling the CCT evolutionary trajectory (Figure 4E), it posed a clear mutation rhythm corresponding to the virus global outbreaks. In contrast, CGA had the most deletion (Fr < -3000) but its trajectory was noise from beginning to the end, indicating its unstable role. Therefore, to become a top leader, a feature must do the right thing at the right time like CCT. Noise-makers like CGA never make difference in evolution.

Discussion

This study developed a novel computational system to reveal the virus evolutionary trajectories. This system computes Fr between any variant sequences versus reference and then employs LSTM to find a clear trajectory, which makes this system robust. Comparing with traditional phylogenetics, this system does not depend on alignment and it catches up the major variations between the sequences, instead of finding the minor variations inside the aligned-similarity as does phylogenetics, but it does not intend to replace traditional phylogenetics method. This system has successfully revealed the evolutionary trajectory of SARS-CoV-2 genome here and the evolutionary trajectory of SARS-CoV-2 variant in my paralleled study¹⁴. Certainly, this Fr-LSTM system can be applied to any evolutionary study, specially the big data study.

Evolutionary trajectory of SARS-CoV-2 genome has been widely regarded as one of the most crucial topics in SARS-CoV-2 studies^{1,8,9,16}, but it has not been revealed. This study systematically uncovered evolutionary trajectories of all possible features (84) in SARS-CoV-2 genome. All features contributes to the virus evolution, but several of them (e.g. CGA) have flexible signals and create noise around the evolutionary trajectory and their functions may be mixed. Only features with clear evolutionary trajectories fitting with virus outbreaks are the most important ones such as CCT and CT. This is why the evolutionary trajectory is so crucial to appreciate the full dynamic picture of virus evolution and its pandemic secret. Results derived from static data may be biased.

Among the 84 genome features, this study found that single nucleotide C and C-prefixed dinucleotides and codons play the central role in driving SARS-CoV-2 evolution. For example, CCT and CT centrally control the virus genome although they have not investigated. In addition, this study also found CG as the biggest content loser in SARS-CoV-2 genome during evolution, paralleling with the

recent study showing CG deficiency in this virus genome¹⁷. Of course, CG is not severely deficient in wild type SARS-CoV-2 (before 3/2021) but it only losses during evolution(after 3/2020 as shown in our data). Therefore, SARS-CoV-2 sounds like a C-virus.

This study uncovers the mystery of SARS-CoV-2 evolution, which will guide the flight against this virus. The novel system established here provides an alternative platform to study any organism evolution.

Materials and methods

General computational environment

All data download, processing, computations and graphing have been done under Linux with python 3.8 and R 3.6. Deep learning neural network was performed by using TensorFlow 2.4.0 and Scikit-learn 0.24.0.

Data resources

All data were downloaded from GISAID (<https://www.gisaid.org/>) on July 4, 2021. Total 2,212,864 genome sequences were downloaded. These samples were subjected to a series of filtering processes, including length (>290k), N content<= 10, and only ATCGN. The entire filtered sequences were split into human group and animal groups. The human group contained total 1,128,954 genome sequences.

Feature selection

This study used 84 genome features, 4 single nucleotides, 16 all possible dinucleotides and 64 codons. Including: 'G', 'C', 'A', 'T', 'CG', 'GC', 'AT', 'TA', 'CA', 'AC', 'CT', 'TC', 'GA', 'AG', 'GT', 'TG', 'TT', 'AA',

'GG', 'CC', 'TTT', 'AAA', 'TTA', 'TGT', 'TTG', 'ACA', 'ATT', 'AAT', 'CTT', 'ATG', 'TAA', 'CAA', 'GTT', 'ACT', 'TGA', 'TAT', 'AAC', 'TAC', 'AGA', 'AAG', 'CTA', 'TGG', 'GTG', 'TCA', 'TGC', 'TCT', 'GAA', 'GCT', 'TTC', 'AGT', 'CTG', 'CAT', 'ATA', 'GTA', 'CAC', 'GGT', 'GAT', 'CAG', 'TAG', 'ACC', 'GCA', 'CCA', 'CCT', 'GAC', 'ATC', 'AGG', 'AGC', 'GAG', 'CTC', 'GGA', 'GTC', 'GGC', 'TCC', 'GCC', 'CGT', 'ACG', 'GGG', 'CCC', 'TCG', 'CGC', 'CGA', 'GCG', 'CGG', 'CCG'.

Discrete Frechet distance (Fr)

This study computed the coupling Fr^{12} for each feature (e.g. “A”) for a given individual virus genome (e.g. EPI_ISL_601443, alpha variant) against the reference genome(NC_045512).

Given P and Q as a feature (e.g. “A”) trajectory for the reference and a variant respectively.

$P = \{p[1], p[2], \dots, p[n]\}$, ordered positions and contents in reference genome

$Q = \{q[1], q[2], \dots, q[m]\}$, ordered positions and contents in variant genome

A coupling L between P and Q is a sequence

$(P[a_1], Q[b_1]), (P[a_2], Q[b_2]), \dots, (P[a_l], Q[b_l])$

where $a_1 = 1, b_1 = 1, a_l = n, b_l = m$. For all $i = 1, \dots, l$, $a_{i+1} = a_i$ or $a_{i+1} = a_i + 1$, and $b_{i+1} = b_i$ or $b_{i+1} = b_i + 1$.

Fr between Q and P is defined as

$Fr(Q, P) = \min\{ \max \text{distance}(Q[b_i], P[a_i]) \text{ for all possible couplings between P and Q} \}$

Frs for all 84 features of more than one million virus were deposited in the project website

Machine learning

Two independent machine-learning sets were prepared in this study, including test (5%) and training (95%) following the order of time series.

The forecasts were set 30 days after July 4, 2021.

All data was pre-processed by MinMaxScaler from klearn.

Deep learning neural network implemented in Keras Sequential library under Tensorflow was used throughout the whole study to predict and forecast. Batch size and epochs were set to 64 and 50 for all machine learning.

To avoid over-fitted, I set dropout (0.2) for two model layer for all models in this study. 4 hidden layers were set to a typical running. Adam was used for model optimizer and mean_squared_error was set for the loss.

Network construction

the associated network was built by FINET¹⁵ by setting stability selection group number =100, frequency cutoff=1, and default for other parameters. Network centrality was calculated by networkx (<https://networkx.org/>).

Final graphing

Several final summary were drawn by using ggplot2 in R. Otherwise, they were completed by python.

All detailed figures and data was deposited in the project website

<https://combai.org/ai/covidgenome/>

Conflict interests

NO

Funding

No funding resource associated with this project

Acknowledgment

Thank GISAID (<https://www.gisaid.org>) for providing the full data.

Figure legends

Figure 1. Computing Fr and evolutionary trajectory. A, principle and algorithm of fr calculation on 3 typical sequence features, including a single nucleotide (A), dinucleotide (AA), and codon (AAA). B, Total 84 feature Fr comparison between three virus groups, Alpha, China, and global. Alpha Fr denotes the Fr of a single virus (EPI_ISL_601443/B117), but China and global Fr represent the Fr median of Chinese samples collected in 2019 and all global samples respectively. This plot followed the order of the Chinese median. C, The Fr median of all 84 features across all daily samples in the UK(United Kingdom). All Fr represents the median of Fr in this study unless specific notice. The LSTM model performed prediction and forecast with MAE and MSE of 0.015358 and 0.000399 respectively. D, The actual infection cases and LSTM prediction cases in the UK (MAE=0.206033 and MSE=0.045503).

Figure 2. The trajectory of 4 single nucleotides, C(A), G(B), A(C),T(D).

Figure 3. The trajectory of dinucleotide. Trajectories of 4 dinucleotides, CG, GC, CT, AT (A-D. E, Maximum absolute Fr predicted for total 16 dinucleotides.

Figure 4. The trajectory of codon. Trajectories of 7 codons, TTA, GGG, ATT, GCT, CCT, CGT, and CAA (A-G). H, Maximum absolute Fr predicted for 64 codons.

Figure 5. Associate network of 84 features. The network (left) and top 10 hubs (right) was computed from all global samples. Yellow nodes denote the first neighbors of CCT(the center of the network). Edge color changes from red (positive interaction) to blue(negative interaction). Edge thickness represents coefficient degree, thicker, higher in coefficient. In the top hub table, the second and third columns were two centrality data, eigenvector and degree, and the final rank was based the two centrality rank score¹⁵.

Reference:

1. Telenti, A. *et al.* After the pandemic: perspectives on the future trajectory of COVID-19. *Nature* **596**, 495–504 (2021).
2. Kraemer, M. U. G. *et al.* Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* **373**, 889–895 (2021).
3. Gonzalez-Reiche, A. S. *et al.* Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* **369**, 297–301 (2020).
4. Faria, N. R. *et al.* Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021).

5. Cohen, J. The hunt for SARS-CoV-2's ancestors heats up. *Science* (2021)
doi:10.1126/science.acx8985.

6. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
7. Wang, A., Thurmond, S., Islas, L., Hui, K. & Hai, R. Zika virus genome biology and molecular pathogenesis. *Emerg. Microbes Infect.* **6**, 1–6 (2017).
8. Singh, D. & Yi, S. V. On the origin and evolution of SARS-CoV-2. *Exp. Mol. Med.* **53**, 537–547 (2021).
9. Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol.* **5**, 1408–1417 (2020).
10. Oude Munnink, B. B. *et al.* The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. *Nat. Med.* **27**, 1518–1524 (2021).
11. Song, K. *et al.* New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Brief. Bioinform.* **15**, 343–353 (2014).
12. Eiter, T. & Mannila, H. Computing Discrete Fréchet Distance, https://openproceedings.org/2019/conf/edbt/EDBT19_paper_219.pdf. 8.
13. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).
14. Wang, A. Evolutionary trajectory and origin of SARS-CoV-2 variant, <https://arxiv.org/abs/2110.07696>. (2021).
15. Wang, A. & Hai, R. FINET: Fast Inferring NETWORK. *BMC Res. Notes* **13**, 521 (2020).
16. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 409–424 (2021).
17. Xia, X. Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense. *Mol. Biol. Evol.* **37**, 2699–2705 (2020).

Fig.2

Fig.3

Fig.4

A TTA

B GGG

C ATT

D GCT

E CCT

F CGT

G CGA

H Predicted 64 codon max fr

Fig.5