Optimizing viral genome subsampling by genetic diversity and temporal distribution (TARDiS) for Phylogenetics

Simone Marini,[1,2], Carla Mavian,[2,3], Alberto Riva,[4], Marco Salemi,[2,3*], and Brittany Rife Magalis.[2,3*]

[1]Department of Epidemiology, University of Florida
[2]Emerging Pathogens Institute, University of Florida
[3]Department of Pathology, University of Florida,
[4]ICBR, University of Florida

*To whom correspondence should be addressed: brittany.rife@epi.ufl.edu, salemi@pathology.ufl.edu

Abstract

TARDiS for Philogenetics is a novel tool for optimal genetic subsampling. It optimizes both genetic diversity and temporal distribution through a genetic algorithm. TARDiS, along with example data sets and a user manual, is available at https://github.com/smarini/tardis-phylogenetics

1 Introduction

Viral genetic sequence data can be used for integrated phylogenetic and population genetic, or phylodynamic, analysis to trace viral evolutionary patterns, as well as spatiotemporal origin and dissemination of viral and bacterial pathogens [5]. Phylodynamic tools, such as NextStrain [6] and BEAST [23, 1], are now routinely utilized to monitor evolution and population dynamics of epidemics based on real-time deposition of pathogen sequences in databases (e.g. GenBank, HIVdatabases, GISAID) [22, 14, 12, 10, 15, 13, 24]. Not unlike traditional epidemiological analysis, however, these methods can significantly be affected by sampling bias [7], and sampling during outbreaks are rarely performed randomly from a representative, stratified population [21]. Not only do the quality and quantity of sequences vary per country, but even regional sample collection policies tend to be inconsistent over time, as exemplified by the inherent sampling bias of SARS-CoV-2 strains, collected through convenience sampling, and

sequenced during the early pandemic phase [15]. Moreover, continuous generation of new sequences can very quickly approach information overload. For example, as of January 15th, 2021, ∼375,000 sequences have been deposited in GISAID (SARS-CoV-2) database, with a number of countries either over or under represented compared of their actual infection prevalence. In such cases, full dataset analyses cannot be accomplished, as computational tools are not designed to handle hundred of thousands of sequences. In order to reduce computational complexity, subsampling must often be performed [8], typically using an approach that maximizes genetic diversity among subpopulations [2] (e.g., countries or regions [8]), which increases phylogenetic signal in the dataset, thus improving phylodynamic inference over convenience sampling. Besides enhancing signal for statistical phylogenetic inference, reliable estimates of significant events in the context of space and time also require sufficient temporal signal in the dataset [7], or distribution of sampling over time, to calibrate reliable molecular clocks [19]. Despite sampling strategies pose a significant threat to conclusions drawn from phylodynamic inference, this problem has received so far insufficient attention [4]. Hall *et al.* (2016) [7] were able to demonstrate that sampling sequences uniformly with respect to both space and time leads to better solutions than optimizing sampling solely by genomic diversity. There currently exists no tool to aid researchers to optimize pathogens' sequences subsampling with respect to space, time, and genetic diversity. In what follows, we introduce TARDiS (Temporal And diveRsity Distribution Sampler), a machine learning approach designed to optimize phylogenetic subsampling according to both genetic diversity and temporal distribution for user-defined subpopulations.

2 Methods

2.1 Genetic Algorithm

TARDiS implements a genetic algorithm (GA) [9, 3] optimizing genetic diversity and time sampling distributions criteria for any set of viral or bacterial genomes. The output consists of user-defined number n of optimally subsampled genomes from a complete dataset of N genomes. Briefly, the algorithm is initialized as a population of random individuals. Each individual is a solution to the problem, i.e., a subsample of size n genomes. Each individual is characterized by a fitness score, reflecting how well that particular individual (solution) performs on the given problem. In our case, fitness is measured as a combination of genetic diversity (i.e., how diverse are the genomes represented by the individual), and time distribution (i.e., how evenly distributed are the genomes represented by the individual along the epidemic timeline).

Genetic diversity maximization. We aim to recover a subsample of genomes as genetically diverse as possible. To do so, we first need to calculate the genetic distance between all possible genome pairs, represented by a square distance

matrix D, with N rows and columns. The user can provide their own distance matrix, or let TARDiS compute it using the Jukes-Cantor substitution model. We calculate the genetic diversity fitness F_{gd} of a subset as

$$F_{gd} = \frac{\sum_{(i,j,i\neq j)}^{n_c} dist(i,j)}{dist_{max}(1,\ldots,n_c)}$$

where n_c are all the genome pairs $(i,j) \in n$, with $i \neq j$; $dist(i,j)$ is the genetic distance of a genome pair (i,j); and $dist_{max}(1,\ldots,n_c)$ is the sum of the genetic diversities of the maximum n elements of the distance matrix D, representing a theoretical upper bound to force fitness $\in [0,1]$, with a higher value representing a better F_{gd}.

Time distribution optimization. Our objective is to recover a subsample of n genomes that are evenly distributed along the considered time interval. Intuitively, if $n = 10$ and the time interval is 10 days, we would like to consider one genome per day. We can thus calculate the ideal time distribution I_{td} as a date vector of n elements, starting with the first available date d_f, ending with the last available date d_l, and having the remaining $n-2$ elements distanced with a $\frac{d_f - d_l}{n-1}$ interval. The worst possible time distribution W_{td}, on the other hand, is a time distribution concentrated into a single specific date (i.e., all samples collected on the same day). We measure the time distribution fitness F_{td} as

$$F_{td} = 1 - \frac{\sum_i^n |time(g_i) - t_i|}{\sum_i^n |t_w - t_i|}$$

where $time(g_i)$ is the collection date of the i-th genome, t_i is i-th date in I_{td}, and t_w is i-th date in W_{td}. In other words, F_{td} is bounded $\in [0,1]$, with a higher value representing a better F_{td}. The final fitness F of a specific individual is calculated as

$$F = F_{gd} \times w_{gd} + F_{td} \times w_{td}$$

where w_{gd} and w_{td} are user-defined weights to set the importance of genetic diversity and time distribution, respectively.

GA operators. Once a population is generated, fitness is calculated for each individual. Individuals are then chosen and combined to produce a new population, in an iterative fashion. To generate a novel individual, TARDiS uses three operators: selection, mutation, and crossover. The selection operator is based on deterministic tournament selection with $k = 5$ [3]. Briefly, two sets of k individuals are randomly chosen, and the individual with the highest fitness is selected from each set. The crossover operator combines two tournament winners A and B into a new individual C by keeping the all the g genomes $\in (A \cup B)$, and randomly selecting $\frac{n-g}{2}$ genomes $\in (A \cap B) - (A \cup B)$. To help avoid local maxima, each newly generated individual C has a 0.08 probability of mutating [9, 3]. A mutation is defined as swapping a genome of individual C with one randomly chosen from the remaining, non-$(A|B)$ genome pool. Note

also that the user defines a fraction of the population that is randomly created (and thus not evolved) for each generation. Another user-defined value determines elitism, i.e., the fraction of best genomes (ranked by fitness) to be copied without modifications in the next generation.

2.2 Case study: subsampling a rising epidemic.

We simulated a growing epidemic using a stochastic, agent-based model [11] with limited migration between ten subpopulations, or regions (a, ..., j). Details on simulation paramaters can be found in the supplementary materials. We ran TARDiS on a single simulated dataset, subsampling 40 genomes per region (with the exception of region f, with 37 genomes available) for 50 generations, with a population of 1000 individuals per generation, of which 85% were evolved, 10% were newly generated, and 5% were elite. The phytools package [20] in R [17] was used for joint likelihood reconstruction of discrete ancestral states [16] according to subpopulation for each internal node of the subsampled trees. Transition rates among discrete states along tree nodes were considered to be equal *a priori*. Migration rates between states were then re-estimated, as described in the supplementary methods. We compared the results obtained both with ($w_g d = 1$, $w_t d = 1$) and without considering time distribution ($w_g d = 1$, $w_t d = 0$). Our simulation indicated that better results are obtained by considering time distribution: the overall migration rate root mean squared error (RMSE) decreased by 17% (0.035 to 0.029). Eight representative clades were then chosen for which the majority of taxa consisted of a single subpopulation and were consistent across true and subsampled trees (Figure 1, A). The RMSE decreased by 43.4% (from 25.37 considering only genetic diversity (t0) to 14.36 if we include time distribution (t1). The addition of a temporal weighting component for an exponentially growing population can act to both increase and decrease representation of earlier time points (e.g., weeks 15 and 12, respectively; Figure 1, B). However, representation of week 1 of the epidemic was increased from 0% to 5%. As the early stages of an epidemic, and time nearing the root of the tree, represent periods of high epidemiological and phylogenetic uncertainty, sample representation during this time is critical for reliable phylodynamic inference and thus contributed to the loss of error in our estimates.

Simulation principles. Each subpopulation was allowed to emerge from the initially infected population (a) with a mean probability of [initial] infection of 0.02 (standard deviation [sd] of 0.005). Each infected individual within a subpopulation was then allowed to migrate to another subpopulation with a mean probability of 0.01 (sd=0.005). The number of contacts for each individual was picked from a normal distribution with a mean of 4 (sd=2). The probability of transmission (when a contact occurs) was provided in the form of a threshold function: prior to 5 days (sd=3), the host was not able to transmit, but after that time, the individual was able transmit with a mean probability of 0.05 (sd=0.005), representing an incubation period for the simulated virus.

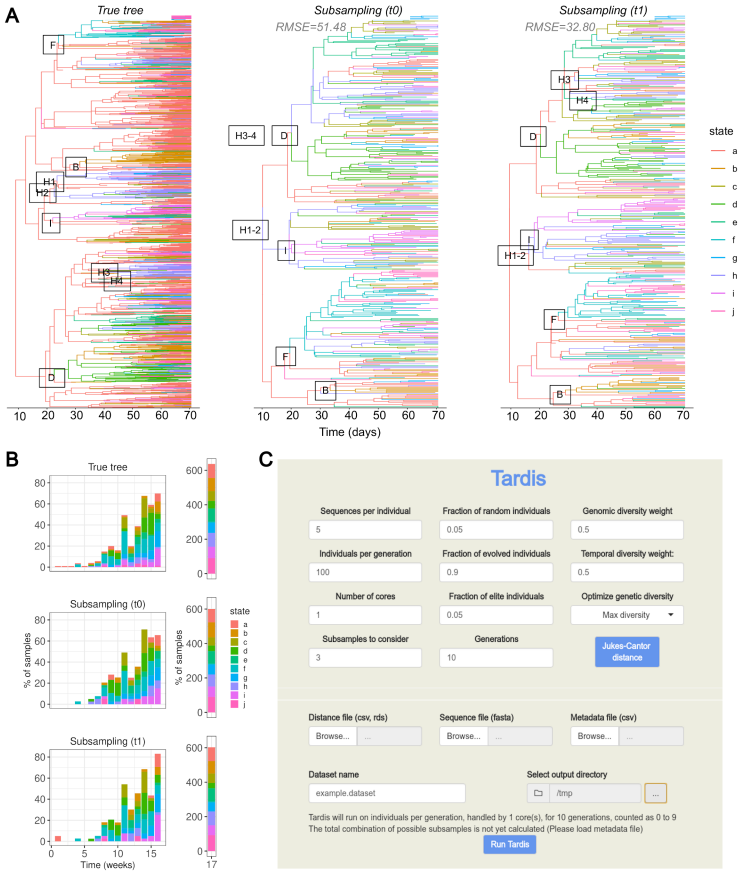

Figure 1: A) True and subsampled trees with representative clades. Eight representative clades were chosen for which the majority of taxa consisted of a single subpopulation, or state, and were consistent across true and subsampled trees. Root mean squared error (RMSE) was calculated for the true times to the most recent common ancestors (TMRCAs) and estimated TMRCAs across the five representative clades for the subsampled tree with (t1) and without consideration of time (t0). (B) Temporal distribution of samples per subpopulation for true and subsampled trees with (t1) and without consideration of time (t0). (C) Screenshot of the TARDiS graphical user interface.

Each infected individual was removed from the simulation (representing death, recovery, etc.) after 14 days. The described parameters resulted in a basic reproductive number (R_0) of approximately 1.6 for the epidemic. The simulated epidemic was run for 365 days or until a total of 10,000 hosts were infected. For each of the ten subpopulations, individuals belonging to that subpopulation were binned according to week of removal (i.e., seven-day intervals) and subsam-

pled according to an exponential distribution (rate=5), representing idealistic sampling of a population proportional to the size of the epidemic and resulting in a range $[37, 844]$ of sampled individuals for each subpopulation (state). The original transmission tree was pruned, leaving only the remaining sampled individuals.

A molecular clock, or constant evolutionary rate across all branches of the tree, was assumed, allowing branches separating noes within the tree to be scaled in both time and genetic distance. Nucleotide (A,C,G,T) sequences were thus simulated along the tree using a general time reversible evolutionary model, with rate matrix (0.32512,1.07402,0.26711,0.25277,2.89976,1.00000) and nucleotide frequencies (0.299,0.183,0.196,0.322). A gamma distributed of rate variation across sites (alpha=2.35) was also used, with a proportion (0.60) of sites considered to be invariable. Branch lengths were scaled by a factor of 8e-04 (representing an approximate evolutionary rate in substitutions/site/year). Sequence simulation was performed in Seq-Gen [18].

Migration rates for each of the ten subpopulations were calculated as a function of the number of transitions between subpopulation states (non-reversible) along each branch within the tree and the frequency (F) of the initial subpopulation among tree tips. I.e, for w branches with transitions between subpopulations, and x branches with specifically transitions from i (node at earlier time point) to j (node at more recent time point),

$$R_{ij} = \frac{x \times F_i}{w}$$

3 Implementation

TARDiS is implemented as a command-line tool based on NextFlow, suitable for analyzing large datasets in an HPC environment, and as a GUI based on R/Shiny for ease-of-use and experimentation 1. TARDiS is available at https://github.com/smarini/tardis-phylogenetics.

References

[1] Remco Bouckaert, Timothy G Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, et al. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650, 2019.

[2] Olga Chernomor, Bui Quang Minh, Félix Forest, Steffen Klaere, Travis Ingram, Monika Henzinger, and Arndt von Haeseler. Split diversity in constrained conservation prioritization using integer linear programming. *Methods in ecology and evolution*, 6(1):83–91, 2015.

[3] Jesús Guillermo Falcón-Cardona and Carlos A Coello Coello. Indicator-based multi-objective evolutionary algorithms: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 53(2):1–35, 2020.

[4] Simon DW Frost, Oliver G Pybus, Julia R Gog, Cecile Viboud, Sebastian Bonhoeffer, and Trevor Bedford. Eight challenges in phylodynamic inference. *Epidemics*, 10:88–92, 2015.

[5] Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James LN Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *science*, 303(5656):327–332, 2004.

[6] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 05 2018.

[7] Matthew D Hall, Mark EJ Woolhouse, and Andrew Rambaut. The effects of sampling strategy on the quality of reconstruction of viral population dynamics using bayesian skyline family coalescent methods: A simulation study. *Virus evolution*, 2(1), 2016.

[8] Samuel L Hong, Simon Dellicour, Bram Vrancken, Marc A Suchard, Michael T Pyne, David R Hillyard, Philippe Lemey, and Guy Baele. In search of covariates of hiv-1 subtype b spread in the united states—a cautionary tale of large-scale bayesian phylogeography. *Viruses*, 12(2):182, 2020.

[9] Oliver Kramer. *Genetic algorithm essentials*, volume 679. Springer, 2017.

[10] John Lednicky, Marco Salemi, Kuttichantran Subramaniam, Thomas B Waltzek, Tara Sabo-Attwood, Julia C Loeb, Shannon Hentschel, Massimo Tagliamonte, Simone Marini, Md Mahbubul Alam, et al. Earliest detection to date of sars-cov-2 in florida: Identification together with influenza virus on the main entry door of a university building, february 2020. 2020.

[11] S. Lequime, P. Bastide, S. Dellicour, P. Lemey, and G. Baele. nosoi: A stochastic agent-based transmission chain simulation framework in r. *Methods Ecol Evol*, 11(8):1002–1007, Aug 2020.

[12] Brittany Rife Magalis, Andrea Ramirez-Mata, Anna Zhukova, Carla Mavian, Simone Marini, Frederic Lemoine, Mattia Prosperi, Olivier Gascuel, and Marco Salemi. Differing impacts of global and regional responses on sars-cov-2 transmission cluster dynamics. *bioRxiv*, 2020.

[13] Carla Mavian, Simone Marini, Mattia Prosperi, and Marco Salemi. A snapshot of sars-cov-2 genome availability up to april 2020 and its implications: Data analysis. *JMIR public health and surveillance*, 6(2):e19170, 2020.

[14] Carla Mavian, Taylor K Paisie, Meer T Alam, Cameron Browne, Valery Madsen Beau De Rochars, Stefano Nembrini, Melanie N Cash, Eric J Nelson, Taj Azarian, Afsar Ali, et al. Toxigenic vibrio cholerae evolution and establishment of reservoirs in aquatic ecosystems. *Proceedings of the National Academy of Sciences*, 117(14):7897–7904, 2020.

[15] Carla Mavian, Sergei Kosakovsky Pond, Simone Marini, Brittany Rife Magalis, Anne-Mieke Vandamme, Simon Dellicour, Samuel V Scarpino, Charlotte Houldcroft, Julian Villabona-Arenas, Taylor K Paisie, et al. Sampling bias and incorrect rooting make phylogenetic network tracing of sars-cov-2 infections unreliable. *Proceedings of the National Academy of Sciences*, 117(23):12522–12523, 2020.

[16] Tal Pupko, Itsik Pe, Ron Shamir, and Dan Graur. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution*, 17(6):890–896, 06 2000.

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[18] Andrew Rambaut and Nicholas C. Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 06 1997.

[19] Andrew Rambaut, Tommy T Lam, Luiz Max Carvalho, and Oliver G Pybus. Exploring the temporal structure of heterochronous sequences using tempest (formerly path-o-gen). *Virus evolution*, 2(1):vew007, 2016.

[20] Liam J. Revell. phytools: An r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3:217–223, 2012.

[21] Brittany D Rife, Carla Mavian, Xinguang Chen, Massimo Ciccozzi, Marco Salemi, Jae Min, and Mattia CF Prosperi. Phylodynamic applications in 21 st century global infectious disease research. *Global Health Research and Policy*, 2(1):13, 2017.

[22] Yuelong Shu and John McCauley. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance*, 22(13):30494, 2017.

[23] Marc A Suchard, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond, and Andrew Rambaut. Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus evolution*, 4(1):vey016, 2018.

[24] Eduan Wilkinson, Dennis Maletich Junqueira, Richard Lessells, Susan Engelbrecht, Gert van Zyl, Tulio de Oliveira, and Marco Salemi. The effect of interventions on the transmission and spread of hiv in south africa: a phylodynamic analysis. *Scientific reports*, 9(1):1–12, 2019.