Modelling Sequence Uncertainty

David Champredon and Art Poon
Western University, London, Ontario, Canada.

2020-07-09

Contents

1 Introduction

Sequencing error. Extracting DNA/RNA from biological samples is a complex process that involves several steps: extraction of the genetic material of interest (avoiding contamination with foreign/unwanted genetic material); reverse transcription (if RNA); DNA fragmentation of the genome into smaller segments; amplification of the fragmented sequences using PCR; sequencing the fragments (*e.g.,* with fluorescent techniques); putting back the small fragments together by aligning them (de novo) or mapping them to benchmark libraries.*(((all this must be checked by someone who knows well the process!)))* Errors can be introduced at each of these steps for various reasons [Beerenwinkel and Zagordi, 2011] and some errors can be quantified (*e.g.,* sequencing quality scores from chromatographs).

In-host diversity and polymorphisms. When the phylogenic tree to infer is based on pathogen sequences infecting hosts, the potential genetic diversity of the infection adds a complexity in phylogeny reconstruction. Typical examples are epidemiological studies reconstructing transmission trees from viral genetic sequences (*e.g.,* HIV, HepC) sampled from infected patients *((ref phyloscanner))*.

Current uncertainty management. The different sources of uncertainty described above impact our observations of the actual genetic sequences. There are standard approaches to deal with identifiable observation errors. Base calls that are ambiguous (from equivocal chromatograph curves or because of genuine polymorphisms) are assigned ambiguity codes (*e.g.,* Y for C or T, R for A or G, etc.). Alignment methods are heuristic methods based on similarity scores that generally do not quantify the uncertainty of alignment.*((double check this is indeed the case for MUSCLE, MAPFT, PRANK, ClustalW))* Methods to reconstruct phylogenies usually leave out the uncertainty complexity and settle for sequences composed of the most frequent nucleotides and/or ignore ambiguity codes.

Propagate and quantify uncertainty. In summary, sources of sequencing observation errors are known and, for a few of them, quantified (quality scores, ambiguity codes). But, to our knowledge, the resulting uncertainty has never been propagated and quantified in a statistical framework for downstream analysis (*e.g.,* alignments, phylogenies inferences). *((Check what BALIphy does, this may be the only example of uncertainty propagation))* In other words, genetic sequences are treated as *certain* quantities.

Here we propose a theoretical framework to represent genetic sequence uncertainty and quantify the impact of uncertainty as it is propagated through methods of phylogeny reconstruction.

2 Methods

In the first part of this section, we propose two simple probabilistic frameworks to represent the uncertainty of genetic sequences observations. The second part describes how those theoretical frameworks can be practically informed from data.

2.1 Probabilistic sequences

Here, we describe two theoretical frameworks to model sequence uncertainty at the *nucleotide level* or at the *sequence level*. In both frameworks, the sequence of nucleotides from a biological sample is not treated as a certain observation, but as a collection of possible sequences.

2.1.1 Nucleotide-level uncertainty

We define probabilistically a nucleotide sequence in a matrix form. For a sequence of length ℓ we can write:

$$\mathcal{S} = \begin{array}{c} \\ \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \\ \texttt{-} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & \ell \end{array} \\ \begin{pmatrix} \mathcal{S}_{A,1} & \mathcal{S}_{A,2} & \dots & \mathcal{S}_{A,\ell} \\ \mathcal{S}_{C,1} & \mathcal{S}_{C,2} & \dots & \mathcal{S}_{C,\ell} \\ \mathcal{S}_{G,1} & \mathcal{S}_{G,2} & \dots & \mathcal{S}_{G,\ell} \\ \mathcal{S}_{T,1} & \mathcal{S}_{T,2} & \dots & \mathcal{S}_{T,\ell} \\ \mathcal{S}_{x,1} & \mathcal{S}_{x,2} & \dots & \mathcal{S}_{x,\ell} \end{pmatrix} \end{array} \tag{1}$$

Each column represents the nucleotide position, each row one of the four nucleotide `A,C,G,T` as well as an empty position "`-`" that symbolizes a genuine deletion (not caused by missing data). Hence, $\mathcal{S}$ is a $5 \times \ell$ matrix. Its elements represent the probability that a nucleotide is at given position:

$$\mathcal{S}_{\texttt{n},j} = \mathbb{P}(\text{nucleotide } \texttt{n} \text{ is at position } j) \tag{2}$$

with the special case for a deletion:

$$\mathcal{S}_{\texttt{-},j} = \mathbb{P}(\text{empty position } j) \tag{3}$$

Note that we have for all $1 \leq j \leq \ell$:

$$\sum_{n \in \{\texttt{A,C,G,T,-}\}} \mathcal{S}_{n,j} = 1 \tag{4}$$

Also, the sequence length is stochastic if $\mathcal{S}_{\texttt{-},i} > 0$ for at least one i. The probability that the sequence has the maximum length ℓ is $\prod_{i=1}^{\ell}(1 - \mathcal{S}_{\texttt{-},i})$. We call the matrix $\mathcal{S}$ the *nucleotide-level probabilistic sequence* of a biological sample. The nucleotide (or deletion) drawn at each position is independent from all the other one, so there are 5^ℓ possible different sequences for a given probabilistic nucleotide sequence.

2.1.2 Sequence-level uncertainty

Out of the 5^ℓ possible sequences, the nucleotide uncertainty may assign a positive probability to sequences that are not biologically possible. As an alternative representation and to reduce the space of possible sequences, let's assume we have enough information (either directly observed from data or simulated) to generate a set of m sequences $\mathcal{B} = (\mathcal{B}_i)_{i \in \{1 \dots m\}}$ of all biologically possible sequences. Note that the $\mathcal{B}_i$ do not have necessarily the same length. The observed genetic sequence, s, is a sample from a specified distribution a:

$$\mathbb{P}(s = \mathcal{B}_i) = a(i) \tag{5}$$

We call the set $\mathcal{B}$ the *sequence-level probabilistic sequence.* Note that, because a is a distribution, we must have $\sum_{i=1}^{m} a(i) = 1$.

2.1.3 Examples

If we have the following nucleotide-level probabilistic sequence:

$$\mathcal{S} = \begin{array}{cc} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{c} \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \\ \texttt{-} \end{array} & \begin{pmatrix} 0.9 & 0.05 & 0.99 & 0 & 0 & 0.6 \\ 0 & 0.8 & 0 & 0 & 0.1 & 0.1 \\ 0.1 & 0.15 & 0 & 0.3 & 0.9 & 0 \\ 0 & 0 & 0.01 & 0.7 & 0 & 0.3 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

then there are $2 \times 3 \times 2^3 \times 3 = 144$ possible sequences. The most likely is the one having the highest nucleotides probabilities: ACATGA with probability 0.2694 ($0.9 \times 0.8 \times 0.99 \times 0.7 \times 0.9 \times 0.6$).

If there is a positive probability of deletion for at least one position, then the sequence has a variable length. Let's take the same example as above, but adding one possible empty position:

$$\mathcal{S} = \begin{array}{cc} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{c} \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \\ \texttt{-} \end{array} & \begin{pmatrix} 0.9 & 0.05 & 0.99 & 0 & 0 & 0.6 \\ 0 & 0.8 & 0 & 0 & 0.1 & 0.1 \\ 0.1 & 0.15 & 0 & 0.2 & 0.9 & 0 \\ 0 & 0 & 0.01 & 0.7 & 0 & 0.3 \\ 0 & 0 & 0 & 0.1 & 0 & 0 \end{pmatrix} \end{array}$$

Like above, there is still a 0.2694 probability that the sequence is ACATGA, but now there is a chance that position 4 is deleted. For example, with probability 0.038 the sequence is ACA-GA.

Below is an example for a sequence-level probabilistic sequence $\mathcal{B}$:

sequence	a
ACATGA	0.60
ACATCA	0.12
AGATCA	0.15
ACAGA	0.05
GCATGA	0.08

Sampling from $\mathcal{B}$, we will have for example ACATCA 12% of the time.

2.1.4 Deletions and insertions

By construction, the nucleotide-level probabilistic sequence must be defined with its longest possible length. Deletions are naturally modelled with our representation but insertions have to be modelled using deletion probability.

Consider the following nucleotide-level probabilistic sequence:

$$\mathcal{S} = \begin{array}{cc} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{c} \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \\ \texttt{-} \end{array} & \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.99 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.01 & 0 & 1 \\ 0 & 0.01 & 0 & 0.99 & 0 & 0 \end{pmatrix} \end{array} \tag{6}$$

The low deletion probability for position 2 is straightforward to interpret: about 1% of the time, nucleotide G at position 2 is deleted. The high deletion probability for position 4 means there is a 1% chance of a T insertion at this position. Table 1 illustrates this.

Table 1: Representation of insertions and deletions from $\mathcal{S}$ defined in (6)

sequence	**frequency**
CGAAT	common, 98% of the time
CAAT	rare (1% frequency) G deletion at position 2,
CGATAT	rare (1% frequency) T insertion at position 4
CATAT	very rare (0.01% frequency) deletion and insertion

The representation of deletions and insertions with a sequence-level probabilistic sequence (not nucleotide-level) is straightforward because in this framework the sequences are explicitly written out, so are their deletions/insertions.

2.2 Probabilistic sequences from data

In this section, we suggest possible methods to inform probabilistic sequences from commonly-used sources of data.

2.2.1 Quality scores from FASTQ files

Fragment sequencing error is an error that is quantified with quality (or "Phred") score attributed to each base call from sequencing instrument. The quality score Q is directly related to the error probability: $\epsilon = 10^{-Q/10}$ [**?**] (where Q typically ranges between 1 and 60). The FASTQ file format is the standard representation for combining sequence and observation error. Hence, the uncertainty associated to the base call is quantified by defining the probability that the observed nucleotide is the correct one:

$$\mathbb{P}(\text{nucleotide} = X \mid \text{observed nucleotide} = X) = 1 - \epsilon \tag{7}$$

Unfortunately, this base-call probability relates to only one *focal* nucleotide and we have no information on the probability for the three other possible nucleotides. Hence, we must make a modelling choice regarding the distribution of the remaining probabilities.

Uniform distribution

As a first simplifying step, we ignore insertions and deletions. Given a base call and its associated quality score at each position, we can assume that the other bases are all equally likely with probability $\epsilon/3$. For example, let's assume the output sequence after fragment sequencing and alignment is ACATG and its associated quality scores are respectively $Q = 60, 30, 50, 10, 40$. The probabilistic sequence is:

$$S = \begin{pmatrix} 1-10^{-6} & 10^{-3}/3 & 1-10^{-5} & 10^{-1}/3 & 10^{-4}/3 \\ 10^{-6}/3 & 1-10^{-3} & 10^{-5}/3 & 10^{-1}/3 & 10^{-4}/3 \\ 10^{-6}/3 & 10^{-3}/3 & 10^{-5}/3 & 10^{-1}/3 & 1-10^{-4} \\ 10^{-6}/3 & 10^{-3}/3 & 10^{-5}/3 & 1-10^{-1} & 10^{-4}/3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{8}$$

Usually, the genetic sequence ACATG would be considered as certain and quality scores discarded. In contrast, within the probabilistic sequence framework the probability the sequence is ACATG is only 0.899 ($= (1-10^{-6}) \times (1-10^{-3}) \times (1-10^{-5}) \times (1-10^{-1}) \times (1-10^{-4})$).

Insertions and deletions ("indels") can be included in the uniform framework. Here, we propose that the nucleotides probabilities are defined conditional on an indel but other models are possible. For a given position, the error probability is $\epsilon = 10^{-Q/10}$ (Q is the quality score) and we assume the probability a deletion happens at this position is d. Conditional on not being deleted, the probability to have the base called is $(1-d)(1-\epsilon)$ and the other three nucleotides can occur with probability $(1-d)\epsilon/3$. Hence, if we assume the base call is A, the column of the nucleotide-level probabilistic sequence for that position is

$$\begin{pmatrix} (1-d)(1-\epsilon) \\ (1-d)\,\epsilon/3 \\ (1-d)\,\epsilon/3 \\ (1-d)\,\epsilon/3 \\ d \end{pmatrix} \tag{9}$$

Multinomial distribution

We can also assume a nucleotide-specific multinomial distribution for the remaining possibilities. For each focal nucleotide observed X, a multinomial distribution $\mathcal{M}_X(\theta)$ can be specified, where θ is the vector of probabilities for the ad-hoc nucleotides. For example, if the observed nucleotide is A and its quality score implies an error probability of $\epsilon = 10^{-4}$, the probabilities that the true nucleotide at that position is actually C, G or T are given by $\mathcal{M}_A(\theta)$ with $\theta_A(C) + \theta_A(G) + \theta_A(T) = \epsilon$ and $\theta_A(X) = p(\text{observed nucleotide} = A \mid \text{true nucleotide} = X)$. We also have to specify the distributions $\mathcal{M}_X$ for $X =$ C, G, T (which will have not necessarily the same probabilities θ). Note that the multinomial case collapses to the uniform one when the elements of θ are all equal.

2.2.2 Absence of uncertainty information (FASTA format)

The observation error estimated by the sampling platform may not always be available and only a character string describes the sequence (FASTA format). In this case, we can model the probabilistic sequence in a similar way as in the FASTQ case, except that additional assumptions must be made to palliate the absence of information regarding the observation error.

The Beta-Uniform model. The observation error probability can be modelled at the sequence level as a Beta distribution:

$$\epsilon \sim \text{Beta}\,(\alpha, \beta) \tag{10}$$

The observation error probability for each focal nucleotide position is then drawn from that distribution. The focal nucleotide being the nucleotide given in the character string representing the sequence (*i.e.*, read in a FASTA file). Then, the probabilities for the other three nucleotides is simply distributed uniformly among them. For example, if the focal nucleotide is A, we have for a given position: $\mathbb{P}(\text{A}) = 1-\epsilon$ and $\mathbb{P}(\text{C}) = \mathbb{P}(\text{G}) = \mathbb{P}(\text{T}) = \epsilon/3$ where ϵ was drawn from the Beta distribution. A deletion could also be modelled as in Equation 9. The Beta distribution is a convenient to model a range of uncertainties from complete uncertainty ($\alpha = \beta = 1$) to near certainty (*e.g.*, $\alpha = 10^3$ and $\beta = 10^{-3}$). Finally, several Beta distributions can be used to reflect sections of the genomes that have different observation error probabilities or polymorphism rate.

The Beta-Multinomial model. Similar to what was done with the FASTQ above, instead of distributing uniformly the remaining probabilities for the non-focal nucleotides, we can use a multinomial instead.

2.2.3 Ambiguity codes

When IUPAC ambiguity codes are produced, we define q as the reliability probability, that is the probability the true nucleotide is among the possibilities given by the ambiguity code. Then we can uniformly distribute q to the possibilities offered by the ambiguity code and $(1-q)$ to the other nucleotides. For example, for a given position, the ambiguity code Y represents either a C or a T and is interpreted as:

$$\begin{pmatrix} (1-q)/2 \\ q/2 \\ (1-q)/2 \\ q/2 \\ 0 \end{pmatrix} \tag{11}$$

Note we could also do a multinomial distribution that distribute q among all the choices offered by the ambiguity code. For simplicity, the special case of a uniform distribution was presented here. Finally, we could consider deletions by including the conditional probability of deletion d as in Equation 9 (but was omitted here for clarity).

2.2.4 Polymorphisms data

Both nucleotide-level probabilistic sequence and sequence-level probabilistic sequence can be generated using error-only non-polymorphic data as well as data from studies investigating polymorphisms. The design of the latter studies may vary but a standard data format they generate can be summarized as follow: the genetic material from several specimens of organisms of interests (e.g., a pathogen infecting a host) is sequenced and all polymorphisms encountered are recorded (after alignment). After alignment, the data can be displayed in a matrix where the columns represent the nucleotide position, the rows represent the nucleotide and deletion, and the matrix elements the number of times the nucleotide was found at that position. If this matrix is normalized column-wise, we obtain the sequence-level probabilistic sequence introduced earlier. An example of such a study, that we'll use to run our simulations, can be found in [Zanini et al., 2015]. *((other similar examples?))*

((Example of studies with full length sequences and their respective frequencies?))

2.2.5 Alignments of short reads (SAM files)

Massive parallel sequencing platforms (*e.g.,* Illumina, Oxford Nanopores, etc.) provide a large number of short reads sequences of the biological sample of interest. The length of those short reads are typically much smaller than the genome sequenced, so they have to be aligned and stitch together in order to re-assemble the full genome sequence. The short reads are typically stored in FASTQ files where the observation error of each nucleotide (estimated by the sequencing platform itself)) is indicated by its Phred score. The alignment and assembly of the short reads is performed by a software (internal to the sequencing platform or not *((check this. Examples?))*) and generates a SAM file *((ref))* that efficiently stores the alignments information. The assembly of the short reads in the SAM file can be represented in as an array where the column are the nucleotide positions. The short reads are "stacked" vertically according to the alignment previously run. The number of short reads stacked for a given nucleotide gives the "coverage" of that position. See Figure 1 for an illustration of this SAM file representation.

We can build a sequence uncertainty model using the information of the SAM representation.

Let's consider a nucleotide at a given position which has a coverage of N short reads (that is a column of the SAM graphical representation). We have N observations for this nucleotide as well as the observation error (available from the FASTQ file of short reads). A simple approach *((and the one usually taken?))* to call the base at that position is the plurality consensus: the base that has the highest frequency is the base called. However, a probabilistic approach estimates the probability that the base is, say, A given the N bases observed at that position, that is $\mathbb{P}(\text{“true” base is A}|\text{observations})$. The observations are a collection of N nucleotides. To simplify the notations, we identify only the number of nucleotides identical to the focal base and lump together the ones that are different. For example if the focal base is A, we count the number n of A nucleotide, hence the number of bases that are different from A is $N-n$. For a given position, the probability that the "true" base is A given that n A and $N-n$ non-A are observed is noted $\mathbb{P}(\texttt{A}|obs:\texttt{A}^n\texttt{X}^{N-n})$ where X represents non-A bases (that is C,G,T and the gap -; the order does not matter).

At a given nucleotide position, we assume the following:

- observations are independent from one another **((double-check this is reasonable))**
- the probability to observe any single nucleotide is 0.25 (*i.e.*, observations not biased)
- the distribution frequency of nucleotide is uniform with probability 0.25

Given those assumptions and some algebra using Bayes' theorem, the probability that the "true" base is A given that n A and $N-n$ non-A are observed is

$$\mathbb{P}(\texttt{A}|obs:\texttt{A}^n\texttt{X}^{N-n}) = \left(1+3^{1-n}\prod_{i=1}^{n}\frac{\epsilon_{A_i}}{1-\epsilon_{A_i}}\prod_{i=1}^{N-n}\left(\frac{1}{\epsilon_{X_i}}-\frac{1}{3}\right)\right)^{-1} \tag{12}$$

where ϵ is the observation error probability associated with the quality score from each observation (obtained from the FASTQ file of the short read).

Using Equation 12, we can calculate the probability for all bases A, C, G, T and gap - and populate the matrix of the nucleotide-level probabilistic sequence, as defined by Equation 1, that is

$$\mathcal{S} = \begin{array}{c} \\ \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \\ \texttt{-} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & \ell \end{array} \\ \begin{pmatrix} \mathbb{P}(\texttt{A}|obs_1) & \mathbb{P}(\texttt{A}|obs_2) & \dots & \mathbb{P}(\texttt{A}|obs_\ell) \\ \mathbb{P}(\texttt{C}|obs_1) & \mathbb{P}(\texttt{C}|obs_2) & \dots & \mathbb{P}(\texttt{C}|obs_\ell) \\ \mathbb{P}(\texttt{G}|obs_1) & \mathbb{P}(\texttt{G}|obs_2) & \dots & \mathbb{P}(\texttt{G}|obs_\ell) \\ \mathbb{P}(\texttt{T}|obs_1) & \mathbb{P}(\texttt{T}|obs_2) & \dots & \mathbb{P}(\texttt{T}|obs_\ell) \\ \mathbb{P}(\texttt{-}|obs_1) & \mathbb{P}(\texttt{-}|obs_2) & \dots & \mathbb{P}(\texttt{-}|obs_\ell) \end{pmatrix} \end{array} \tag{13}$$

where obs_i represents the N_i nucleotides observed at position i of the aligned short reads (*i.e.*, N_i is the coverage for position i).

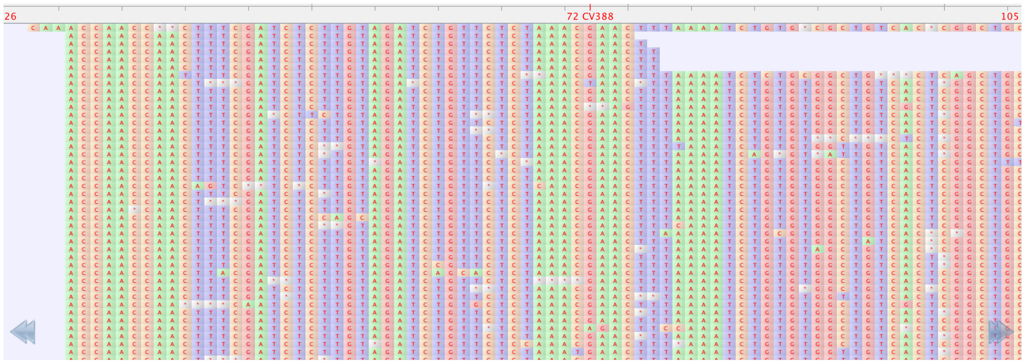

*Figure 1: **SAM file graphical representation.** The software Tablet ((ref)) was used. The 72th nucleotide in this alignment has a coverage of 388...blabla*

2.2.6 Raw acquisition data

Although not commonly available, raw acquisition data files can be used to construct probabilistic sequences. Those files usually contain nucleotide signal intensities as they are read along the DNA/RNA molecular chain. For example, Sanger sequencing generates chromatograms (AB1/I files) from variations in fluorescence while nanopores technologies generate electrical resistance signals (FAST5 files). Nucleotides are "called" from interpreting peaks in the signals. For all technologies, the determination of the nucleotide read from the signal relies on complex and imperfect algorithms (*e.g.,* neural networks for nanopore technologies). Rather than a dichotomous translation of the signal into one single nucleotide (potentially associated with an observation error probability using a quality score), it may be beneficial to take a probabilistic approach in calling the bases from raw signal data.

3 Examples

3.1 Propagating sequence uncertainty in phylogeny reconstruction

Molecular phylogenies are tree-based models that relate common ancestors of genetic sequences. Many sophisticated statistical tools exist to reconstruct phylogenies from genetic material extracted from biological samples. Those statistical methods rely, to a varying degree, on "truthful" and accurate observations of molecular sequences, their main – if not unique – input data.

Here, we describe our study design to propagate and measure sequence uncertainty in phylogeny reconstruction.

3.1.1 Generating simulated probabilistic sequences

If we want to simulate realistic probabilistic sequence, we have to reproduce a similar uncertainty as the one we would have from either sequencing error or polymorphism.

We illustrate our methodology in the context of in-host HIV infections. The data from Zanini [Zanini et al., 2015] is a good source to assess primarily the diversity of polymorphism for HIV, and to a certain extent too, the sequencing error (because it is always here). Briefly, this data set gives, for several patients at several time points during their (untreated) infection, the number of times nucleotides were sample at a given position, across the whole HIV genome. The number of nucleotide occurrences at each position can easily be transformed into the probabilities for the probabilistic sequence. The entropy can then be calculated at each position, and also for the entire genome (by simply summing up the entropies for all positions).

Entropy is a measure of uncertainty. So we can consider the distribution of entropies (for each position on the genome) as a representation of the overall genome sequencing uncertainty, that should be approximately matched by simulations deemed realistic. The data from Zanini and colleagues [Zanini et al., 2015] shows that $\mathcal{S}_{n,j}$, the distribution of base-call probabilities for most positions is highly concentrated just under 1 (which means a high base-call probability for most positions). Hence, we choose a Beta distribution to simulate base-call probabilities, and fit the shape parameters α and β on the observed entropy distribution:

$$S_{n,j} \sim \mathrm{Beta}\,(\alpha, \beta) \tag{14}$$

$$\alpha, \beta \text{ such that } E(\alpha, \beta) = E_{obs} \tag{15}$$

where E is the distribution of position-wise entropy. A fit on Zanini's data [Zanini et al., 2015] gives approximately $\hat{\alpha} = 29.7$ and $\hat{\beta} = 0.06$. *((make an appendix to show the details of this fit.))*

We calculate the entropy value as

$$E(\alpha, \beta) = -\sum_{i=1}^{\ell} p_i \log_2(p_i) \tag{16}$$

where p_i is the (α- and β-dependent) base-call probability drawn for the nucleotide at position i and ℓ is the length of the sequence.

3.1.2 Assessing the impact of sequencing uncertainty

Below is our simulation design to study the impact of uncertainty on phylogeny reconstruction. An illustration of this pipeline is given by Figure 2.

0. Choose a root sequence of interest (*e.g.,* a HIV genome, a random sequence)

1. Generate a phylogeny from this root sequence, using phyloSim. The resulting tree T^* has n tips that represent the sequenced samples $seq_1, seq_2, \dots, seq_n$. The tree T^* with its sequences seq_i is the "base" phylogeny.
2. Add a simulated layer of uncertainty by transforming the "base" sequences seq_i into probabilistic sequences $\mathcal{S}^i$ (for $i = 1, 2, \dots, n$).
3. Repeat M times: draw a sequence $\widetilde{seq_i}$ for each $\mathcal{S}^i$ (for $i = 1, \dots, n$).
4. Repeat M times: reconstruct the phylogeny T_m with RAxML from the $(\widetilde{seq_i})_{i=1\dots n}$.
5. Assess the uncertainty by considering the variance among the phylogenies $(T_m)_{m=1:M}$ using several distance metrics (detailed below).

Note that the M iterations amounts to a Monte-Carlo algorithm. Studying the distance between the reconstructed trees $(T_m)_{m=1:M}$ and the true tree T^* is not our main goal (this distance essentially assesses the performance of the phylogeny reconstruction software to correctly infer the "true" ancestry). Instead, we are principally interested in *uncertainty propagation*, that is the variance of the pairwise distances between the $(T_m)_{m=1:M}$.

Our analysis considers five levels of uncertainty. We start with a virtually inexistent sequence uncertainty, then increased it by lowering the base call probability. This is done by sampling the probability from multiple parameter sets (α, β) of a Beta distribution (see Equation 14). We choose a single value $\alpha = 29$ and use five different values for the second shape parameter $\beta = 10^{-3}, 10^{-2}, 10^{-1}, 1$ and 3 *((update if necessary))*. With these values, the mean of the Beta distribution for the base-call probability decreases away from 1.0. Finally, note that the middle value ($\alpha = 29, \beta = 10^{-1}$) is close to the fitted entropy values of the longitudinal HIV dataset from Zanini and colleagues [Zanini et al., 2015].

For Step 5, we explore the impact of sequence uncertainty on several types of downstream analysis on reconstructed phylogenies: pairwise distance between trees, clustering and an example of source attribution *((amend if needed))*.

Pairwise distances between trees. Define the set

$$D = \{d(T_i, T_j);\ i = 1, \dots, M \text{ and } j < i\} \tag{17}$$

with d a tree distance. The distance d should be a statistically-convenient metric that represents faithfully the differences of interpretation (*i.e.*, uncertainty) of phylogeny reconstruction. We use three distances: Robinson-Foulds (RF) [Robinson and Foulds, 1981], kernel [Poon et al., 2013] and a label-based distance [**?**].

We measure the uncertainty of phylogenetic inference with the coefficient of variation $c = s/m$ where m is the mean of D and s its standard deviation. We note c_{RF}, c_K and c_L the coefficients of variation calculated with the RF, kernel and label-based distances, respectively.

Although not our primary objective in this study, we also investigate the distance of the inferred tree T_i to the benchmark tree T^*, and define

$$D^* = \{d(T_i, T^*);\ i = 1, \dots, M\} \tag{18}$$

Similarly as for c, we define c^* as the coefficient of variation of D^* and adopt the same subscript notation to differentiate between the distances used for its calculation.

Impact on clustering. ((TODO))

Impact on source attribution. ((TODO))

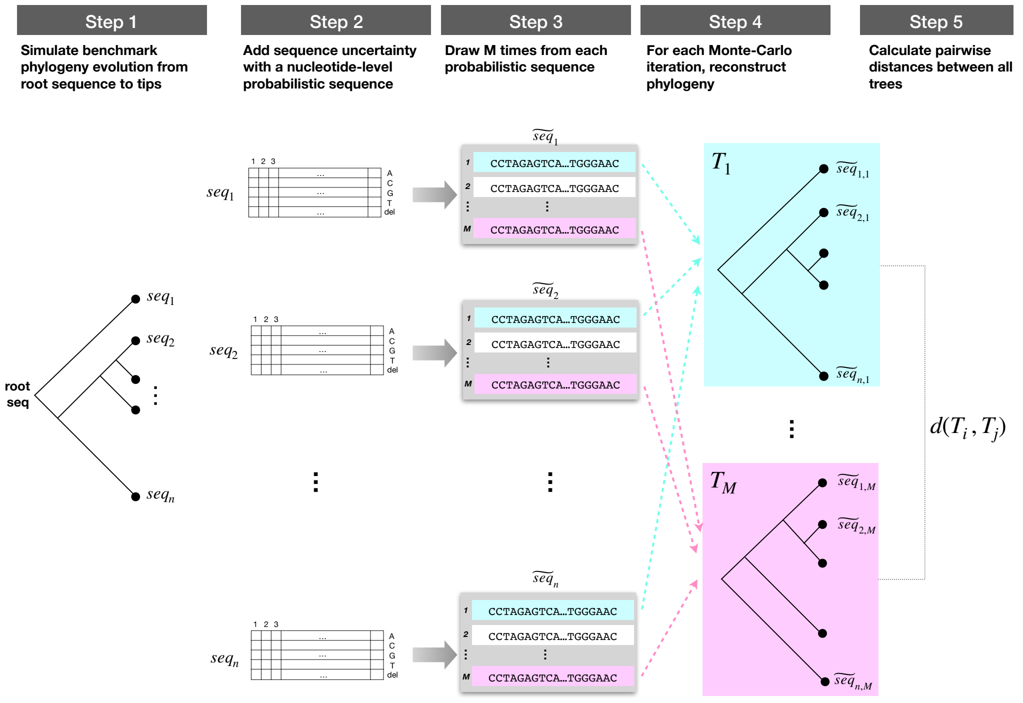

Figure 2: ***Simulations pipeline.*** *Step 1: A phylogeny with n final nodes is simulated from a root sequence using* phylosim*. Step 2: A nucleotide-level probabilistic sequence is generated for each sequence, assuming a Beta distribution for the base-call probability Step 3: For each nucleotide-level probabilistic sequence, a sequence is drawn M times Step 4: Using the ith drawn sequence (i.e., ith Monte Carlo iteration), the phylogeny T_i is inferred ($i = 1, \dots, M$). Step 5: The pairwise distances $d(T_i, t_j)$ are calculated for all $i < j$. Steps 1 to 5 are repeated for several level of uncertainty (defined by the Beta parameters of the base-call probabilities).*

4 Results

References

[Beerenwinkel and Zagordi, 2011] Beerenwinkel, N. and Zagordi, O. (2011). Ultra-deep sequencing for the analysis of viral populations. *Current Opinion in Virology*, 1(5):413–418.

[Poon et al., 2013] Poon, A. F. Y., Walker, L. W., Murray, H., McCloskey, R. M., Harrigan, P. R., and Liang, R. H. (2013). Mapping the shapes of phylogenetic trees from human and zoonotic rna viruses. *PLOS ONE*, 8(11):1–11.

[Robinson and Foulds, 1981] Robinson, D. and Foulds, L. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131 – 147.

[Zanini et al., 2015] Zanini, F., Brodin, J., Thebo, L., Lanz, C., Bratt, G., Albert, J., and Neher, R. A. (2015). Population genomics of intrapatient hiv-1 evolution. *Elife*, 4:e11282.