Propagating Sequence Uncertainty in Phylogeny Reconstruction

Champredon, David Poon, Art

January 22, 2020

1 Introduction

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.

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 in 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

2.1 Probabilistic sequences

Here, we propose two simple probabilistic frameworks to represent the uncertainty of our genetic sequences observations. The first framework represents uncertainty at the *nucleotide level*, whereas the second one is at the *sequence level*. In both frameworks, the sequence of nucleotides from a biological sample is not treated as a certain observation anymore, 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:

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

Each column represents the nucleotide position, each row one of the four nucleotide $\texttt{A,C,G,T}$ as well as an empty position "$\texttt{-}$" 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:

$$\mathcal{S}_{n,j} = \Pr(\text{nucleotide } \texttt{n} \text{ is at position } j) \tag{1}$$

with the special case for a deletion:

$$\mathcal{S}_{\texttt{-},j} = \Pr(\text{empty position } j) \tag{2}$$

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

$$\sum_{n \in \{\texttt{A,C,G,T,-}\}} \mathcal{S}_{n,j} = 1 \tag{3}$$

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 $\prod_{i=1}^{\ell}(1 - \mathcal{S}_{x,i})$. 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:

$$\Pr(s = \mathcal{B}_i) = a(i) \tag{4}$$

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.

Let's take the following 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.

2.2 Probabilistic sequences from data

Here, we suggest possible methods to populate values in probabilistic sequences from data.

2.2.1 Quality scores

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: $\epsilon = 10^{-Q/10}$ [?] (for the widespread Illumina instruments, the sequencing error probability ranges between $10^{-3.5}$ and $10^{-1.5}$ [?]). So each base call is right with probability $1 - \epsilon$. Assuming the other bases and deletion - are all equally likely with probability $\epsilon/4$. Alternatively, if we know only mutations (not deletions) affect the sequence, the last row can be filled with zeros and the other base-calls equal to $\epsilon/3$.

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 can be defined as:

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

Usually, this output from the sequencing instrument would be considered as certain (and quality scores discarded). In the probabilistic sequence framework, the probability to have ACATG is 0.899 ($= (1-10^{-6}) \times (1-10^{-3}) \times (1-10^{-5}) \times (1-10^{-1}) \times (1-10^{-4})$).

2.2.2 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 nucleotide-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?))

2.3 Propagating sequence uncertainty in phylogeny reconstruction

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

2.3.1 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 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:

$$S_{n,j} \sim \text{Beta}\,(\alpha, \beta) \tag{5}$$

$$\alpha, \beta \text{ such that } E(\alpha, \beta) = E_{obs} \tag{6}$$

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

$$E(\alpha, \beta) = -\sum_{i=1}^{\ell} p_i \log_2(p_i) \tag{7}$$

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.

2.3.2 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 1.

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 $seq_1, seq_2, \ldots, seq_n$. 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 $i = 1, 2, \ldots, n$).

3. Repeat M times: draw a sequence $\widetilde{seq_i}$ for each $\mathcal{S}^i$ (for $i = 1, \ldots, n$).

4. Repeat M times: reconstruct the phylogeny T_m with RAxML from the $(\widetilde{seq_i})_{i=1 \ldots n}$.

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 $(T_m)_{m=1:M}$ and 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 5). 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, 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

$$D = \{d(T_i, T_j);\ i = 1, \ldots, M \text{ and } j < i\} \tag{8}$$

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

$$D^* = \{d(T_i, T^*);\ i = 1, \ldots, M\} \tag{9}$$

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.

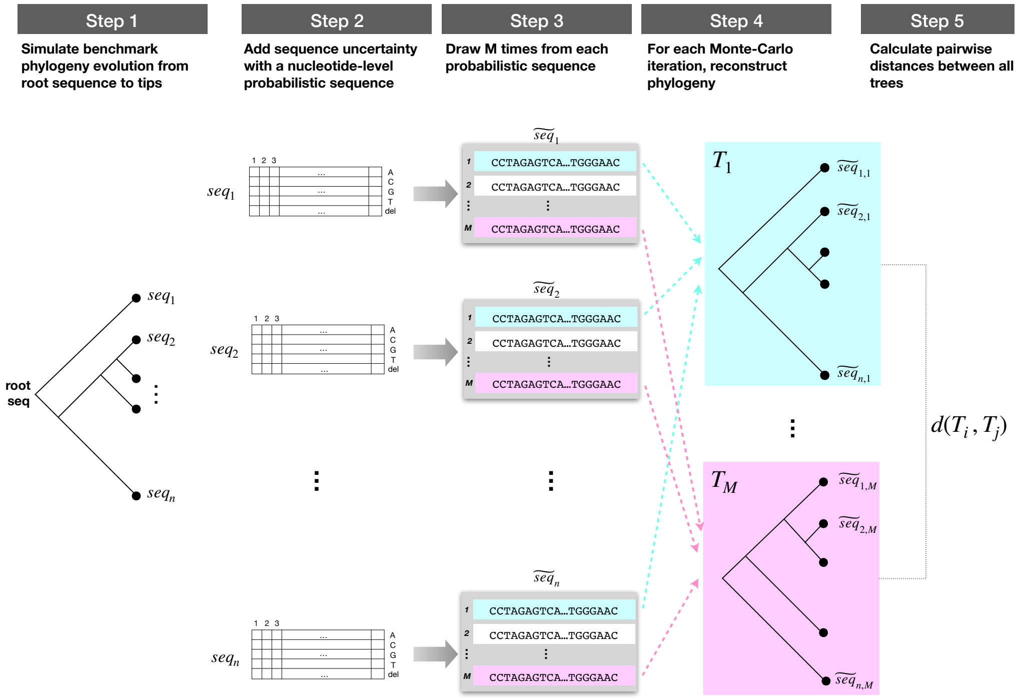

Figure 1: **Simulations pipeline.** *Step 1*: A phylogeny with n final nodes is simulated from a root sequence using phylosim. *Step 2*: A nucleotide-level probabilistic sequence is generated for each sequence, assuming a Beta distribution for the base-call probability *Step 3*: For each nucleotide-level 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 ($i = 1, \ldots, M$). *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).

Impact on clustering. ((TODO))
Impact on source attribution. ((TODO))

3 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.