Propagating Sequencing Uncertainty in Phylogeny Reconstruction

Champredon, David   Poon, Art

December 5, 2019

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!)))* It is well known that errors can be introduced at each of these steps for various reasons and errors can be quantified for some of them (*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 example are epidemiological studies reconstructing transmission trees from viral genetic sequences (*e.g.,* HIV, HepC) sampled from infected patients.

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.). *((Is there uncertainty quantification for alignment methods??))* 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. 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:

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

Each column represents the nucleotide position, each row one of the four nucleotide `A,C,G,T` as well as an empty position `x` that symbolizes a deletion. 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:

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

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

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

Also, the sequence length is stochastic if $\mathcal{S}_{x,i} > 0$ 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 likely assigns a positive probability to sequences that are not biologically possible. As an alternative representation and to reduce the space of sequences, let's assume we have enough information to generate the set $\mathcal{B} = (\mathcal{B}_i)_{i \in \{1 \dots M\}}$ 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*.

2.1.3 Examples

If we have the following nucleotide-level probabilistic sequence:

$$\mathcal{S} = \begin{pmatrix} 0.9 & 0.05 & 0.99 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0.1 \\ 0.1 & 0.15 & 0 & 0.3 & 0.9 \\ 0 & 0 & 0.01 & 0.7 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

then there are $2 \times 3 \times 2^3 = 48$ possible sequences. The most likely is the one having the highest nucleotides probabilities: `ACATG` with probability 0.449 ($0.9 \times 0.8 \times 0.99 \times 0.7 \times 0.9$).

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:

$$\mathcal{S} = \begin{pmatrix} 0.9 & 0.05 & 0.99 & 0 & 0 \\ 0 & 0.8 & 0 & 0 & 0.1 \\ 0.1 & 0.15 & 0 & 0.2 & 0.9 \\ 0 & 0 & 0.01 & 0.7 & 0 \\ 0 & 0 & 0 & 0.1 & 0 \end{pmatrix}$$

Like above, there is still a 0.449 probability that the sequence is `ACATG`, but with probability 0.064 position 4 is deleted and the sequence would be `ACAG`.

Let's take the following example for a sequence-level probabilistic sequence $\mathcal{B}$:

sequence	a
`ACATG`	0.60
`ACATC`	0.12
`AGATC`	0.15
`ACAG`	0.05
`GCATG`	0.08

Sampling from $\mathcal{B}$, we will have for example `ACATC` 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 `x` 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 from polymorphisms data where the distribution a is the abundance of each polymorphism.

((STOPPED HERE. CONTINUE WITH ZANINI'S DATA))

2.3 Distances between phylogenies

((a brief review of the some distances because we'll them in the next section))

2.4 Propagating sequence uncertainty in phylogeny reconstruction

3 Results