Propagating Sequencing Uncertainty in Phylogeny Reconstruction

Champredon, David Poon, Art

December 3, 2019

1 Introduction

Molecular phylogenies are tree-based models of how of genetic sequences are related by common ancestors. For nearly three decades, scientists have developed sophisticated statistical tools 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 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

2.2 Simulating realistic probabilistic sequences

3 Results