Phylogenetics

SubRecon: Ancestral Reconstruction of Amino Acid Substitutions Along a Branch in a Phylogeny

Christopher Monit1,* and Richard A. Goldstein1

1Division of Infection and Immunity, University College London, London, WC1E 6BT, United Kingdom

*To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Existing ancestral sequence reconstruction techniques are ill-suited to investigating substitutions on a single branch of interest. We present SubRecon, an implementation of a hybrid technique integrating joint and marginal reconstruction for protein sequence data. SubRecon calculates the joint probability of states at adjacent internal nodes in a phylogeny, i.e. how the state has changed along a branch. This does not condition on states at other internal nodes and includes site rate variation. Simulation experiments show the technique to be accurate and powerful. SubRecon has a user-friendly command line interface and produces concise output that is intuitive yet suitable for subsequent parsing in an automated pipeline. Availability and Implementation: SubRecon is platform independent, requiring Java v1.8 or above. Source code, installation instructions and an example dataset are freely available under the Apache 2.0 license at https://github.com/chrismonit/SubRecon. Contact: c.monit.12@ucl.ac.uk

1 Introduction

An evolutionary biologist may notice that taxa within a single clade in their sequence dataset possess a distinctive characteristic, such as a unique function. They may wish to investigate the evolutionary events occurring on the ancestral branch dividing this clade from other nodes in the phylogeny, by determining how the ancestral states changed between the two nodes on either side of that branch.

Two ancestral reconstruction techniques are widely used and address distinct statistical questions. Joint reconstruction estimates the set of character states for all internal nodes, reconstructing the whole history of states in the phylogeny (Yang et al., 1995; Pupko et al., 2000). Marginal reconstruction estimates states at a single internal node of interest, without conditioning on states at other internal nodes (Yang et al., 1995; Koshi and Goldstein, 1996). Marginal reconstructions of states at two adjacent nodes will not provide a valid indication of the changes that occurred along the branch connecting them, as the independently estimated states may be incompatible. A complete joint reconstruction provides estimates conditional on the states of all other nodes of the tree, biasing the reconstruction at the nodes of interest.

We have developed a hybrid technique that overcomes these limitations by jointly reconstructing states at nodes either side of a single branch, while marginalising over states at other internal nodes. We present a convenient implementation. SubRecon, which performs this reconstruction for amino acid states. SubRecon is simple to both install and run, has intuitive, configurable output and is suitable for large datasets.

2 Methods

2.1 Theory

We model sequence evolution as a site-independent, time-continuous, reversible Markov process (see, e.g., Yang, 2014). Our approach is applicable to nucleotide, codon or amino acid states, but our implementation considers the latter only. For a given alignment site, we calculate the joint probability of a pair of states at the internal nodes either side of a branch of interest, while marginalising over states at other internal nodes in the phylogeny. This is conditional on states observed at the tip nodes (data, D), a known or estimated phylogeny topology and a substitution rate matrix, Q, with state equilibrium frequencies defined empirically or estimated previously.

Let A and B be the internal nodes connected by the branch, which is of length t, and let a and b represent possible states at these nodes. We use P(t)_{ab} for the a to b transition probability along the branch, where P(t) = exp(Qt). Let πx be the equilibrium probability of state x and P(D|X; θ) the probability of the data D, for nodes descending from node X, conditional on model parameters θ and state x at X, while marginalising over states at other internal descendant nodes, computed with the pruning algorithm (Felsenstein, 1981). We position the root node on the branch of interest, meaning the sets of taxa descending from A and B are mutually exclusive and together comprise all taxa in the phylogeny.

© The Author(s) 2018. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
The joint probability that states $a$ and $b$ existed at $A$ and $B$ respectively, marginalising over other nodes, is thus

$$P(A=a, B=b|D, \theta) = \sum_{a', b'} \pi_A P(t_{ab}) P(D_A|a, \theta) P(D_B|b, \theta)$$

(1)

The root position and designations of $A$ and $B$ are arbitrary since the process is reversible: $\pi_A P(t_{ab}) = \pi_B P(t_{ba})$. The denominator is equal to the marginal probability of the data given the model, i.e., the likelihood, $P(D|\theta)$. The $a$ and $b$ pair maximising $P(A=a, B=b|D, \theta)$ is preferred.

Simulation experiments using various phylogeny topologies, branch lengths and minimum probability thresholds show reconstruction estimates to be accurate and powerful. For mid-sized datasets (fig. 1A-C) the max[$P(A=a, B=b|D, \theta, \alpha)$] ≥ 0.9 threshold yielded between 0 and at most 15 inaccurate reconstructions out of 1000, while even the 0.7 threshold provided a reasonable tradeoff. For very large, highly divergent datasets where the branch of interest is distant from terminal taxa (as in fig. 1D), high minimum thresholds are advisable.

2.3 Software Implementation

SubRecon computes $P(A=a, B=b|D, \theta, \alpha)$ for all $a$ and $b$ pairs, for a specified pair of adjacent internal nodes, using any of several amino acid empirical substitution models; e.g., WAG (Whelan and Goldman, 2001), implemented in PAL (Drummond and Strimmer, 2001). The phylogeny, including branch lengths, and gamma distribution shape parameter $\alpha$ should be estimated in advance using popular phylogeny estimation tools, such as RAxML (Stamatakis, 2014). The models’ default equilibrium frequencies ($\pi$) can be used or estimated values provided.

SubRecon is designed to handle large datasets, as multiple sites can be analysed in parallel with a user-defined number of computing threads and log-transformations prevent numerical underflow errors.

Written in Java v1.8, SubRecon is platform independent and we include build scripts allowing easy compilation using Apache Ant (http://ant.apache.org). Its command line interface (based on JCommander, http://jcommander.org) is simple and the output is intuitive yet amenable to parsing by downstream software in an analysis pipeline. The detail and formatting of output can be controlled by the user.

3 Conclusion

Existing reconstruction implementations are not well suited to comparing ancestral states underlying phylogenetically and biologically distinct taxa in a protein sequence dataset. Our technique combines joint and marginal reconstruction approaches, allowing efficient and valid comparisons. Our convenient implementation, SubRecon, should be a useful addition to the toolkit of the investigator studying comparative evolutionary biology.