Journal of Statistical Planning and Inference 161 (2015) 73–80

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Consistency of a phylogenetic tree maximum likelihood estimator

Arindam RoyChoudhury [a], Amy Willis [b,*], John Bunge [b]

[a] Department of Biostatistics, Columbia University, New York, NY, USA
[b] Department of Statistical Science, Cornell University, Ithaca, NY, USA

ARTICLE INFO

Article history:
Received 28 August 2014
Received in revised form 12 January 2015
Accepted 12 January 2015
Available online 22 January 2015

Keywords:
Phylogenetic tree
Maximum likelihood
Divergent evolution
Parametric models
Asymptotics

ABSTRACT

Phylogenetic trees represent the order and extent of genetic divergence of a fixed collection of organisms. Order of divergence is represented via the tree structure, and extent of divergence by the branch lengths. Both the tree's structure and branch lengths are unknown parameters and the tree is estimated using sequence information sampled at a number of genetic sites. Under the model of genetic Brownian motion, we prove that as the number of genetic sites that are sampled becomes large, the maximum likelihood estimator of the tree is consistent. (Our maximum likelihood estimator treats each site as an independent data point, which is different from concatenating the sites.) Existing arguments for consistency rely on the assumption of a finite parameter space or only apply to transition probability matrix-based models, and do not hold here due to the continuous model for branch lengths. The metric space of Billera et al. (2001) is central to the proof. We conclude with some comments on the role of parametric methods in tree estimation.

1. Introduction

The extent and order of divergence between species is a central topic of research in evolutionary biology. Biologists wish to trace distinct species to a common ancestor, and to model when and in what order evolutionary paths diverged. It is natural to encode this information in a tree, with vertices representing divergence events and branch lengths denoting time between divergence events. Both the order of divergence events (tree structure and leaf labels) and branch lengths (inter-divergence times) are unknown parameters to be estimated.

The frequentist parametric approach to estimation, in particular maximum likelihood, is classical in statistics, but there is much debate in the biological literature about its efficacy. Both nonparametric and Bayesian methods have been found to give viable alternatives for tree estimation (Bouckaert et al., 2014; Ronquist et al., 2012; Alfaro et al., 2003). However, maximum likelihood estimation has well-known asymptotic optimality properties, and consistency is of particular importance in this problem (Rogers, 1997; Yang, 1994). Much attention in statistics has been devoted to finding sufficient conditions for consistency of a maximum likelihood estimator (MLE). These conditions hold in many circumstances but may be hard to verify, or even fail, in nonstandard problems; indeed, examples of seemingly straightforward estimation problems with inconsistent MLEs abound in the statistical literature. Arguments for the consistency of phylogenetic tree MLEs have been put forward (Felsenstein, 1973, 2004; Rogers, 1997; Yang, 1994; Chang, 1996), but they are insufficient for our case, and for this reason we undertake a formal proof of consistency here.

* Corresponding author.
E-mail address: adw96@cornell.edu (A. Willis).

http://dx.doi.org/10.1016/j.jspi.2015.01.001
0378-3758/

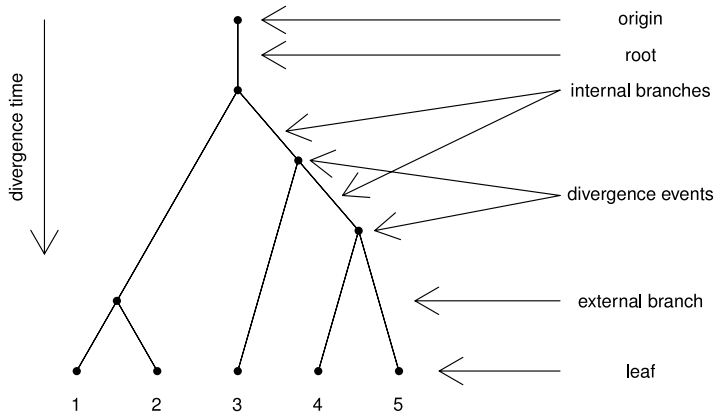

Fig. 1. An example of a phylogenetic tree with 5 leaves, showing the root, internal and external branches, vertices, and the "time arrow". We allow the root to have either zero or positive length.

We first describe the data structure and a general branch length-based model, and discuss the existing literature on tree MLE consistency (Section 2) when each genetic site is considered an independent data-point (as opposed to concatenating the sites). We then describe the shortcomings of existing arguments (Section 3). In Section 4 we formalize the problem statement and the required metric space, and we present a proof of consistency of the tree MLE under the model of Brownian motion along the tree (Section 5). We conclude with some perspectives on the method of proof and on the role of likelihood methods in phylogenetic research (Section 6).

2. A phylogenetic tree model

A phylogenetic tree is a weighted tree-graph which represents biologists' beliefs about the evolution of organisms as characterized by their genetic information. The underpinning belief is that the organisms of interest all evolved from a common ancestor, and due to stochastic changes in genetic information, the ancestral lineage "splits" and creates new organisms as time progresses. Vertices represent organisms that existed at some historical point, and vertices with one edge represent either the common ancestor (the *origin*) or "current" organisms (*leaves*). In our model we allow the root, that is, the distance from the origin to the first split, to be either positive or zero. This is because some researchers regard the root as biologically noninformative, and in this case it may be taken to be zero, that is, omitted from the model. It is assumed that no historical data is available and only information about current organisms may be used to construct the tree. Organisms which share much genetic information are assumed to have diverged later than those with very different information, and the extent of this difference is encoded in the branch lengths. For instance, in Fig. 1 organisms 1 and 2 diverged later than organisms 4 and 5. We wish to estimate the tree structure (including the leaf labels) and the branch lengths based on genetic information from P organisms (in Fig. 1, $P = 5$). Since usually only a finite number of populations are under study, P is fixed throughout.

Since sampling the genetic sites exhaustively is infeasible in practice, we suppose that only n genetic sites are sampled. The simplest form of genetic information is a single base at a site: A, T, C or G, each of which represents a genetic "category". Thus we observe one of four categories for each of P individuals, at n sites. We denote the set of all possible combinations at each site as $\{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}^P$.

The base observed at a given site for one individual is not independent of the same site for another individual, since individuals closer on the tree are more likely to share genetic information (we use this dependence structure to estimate our tree). However, we assume that sites are observed independently. We also assume that the probability distribution for patterns is the same across all sites; that is, the probability of observing a given sequence, say ATCCA (with $P = 5$), does not depend on the site index j. Using only these assumptions, we may write the joint likelihood for the sequences as

$$\begin{aligned} L_{(a_1,\ldots,a_n)} &:= \mathbb{P}(\text{observe sequence } a_1 \text{ at site 1}, a_2 \text{ at site 2}, \ldots, a_n \text{ at site } n) \\ &= \prod_{j=1}^{n} \mathbb{P}(\text{sequence at site } j \text{ is } a_j) = \prod_{j=1}^{n} \mathbb{P}(\text{sequence } a_j). \end{aligned}$$

If instead of indexing the n sequences $a_1, \ldots, a_n$, we index by m unique sequences, then letting the number of sites at which the ith unique sequence is observed be denoted n_i, we may write the log-likelihood as

$$l = \sum_{i=1}^{m} n_i \log p_i,$$

where p_i is the probability of observing sequence i.