# Consistency of the MAP of phylogenetic trees

MATH 833 - Fall 2012 *Presenter: Claudia Solis Lemus*

Reference: [1].

The motivation of the present paper is to provide a formal proof that the Bayesian inference for inferring phylogenetic tree topology from aligned DNA sequence data is statistically consistent under the same identifiability conditions for the consistency of the Maximum Likelihood estimator.

# 1 Setup

## 1.1 Statistical notation

The problem is to identify a certain parameter $a$ from a sequence of iid observations. We will consider two parameters $(a, \theta)$: $a \in A$ discrete parameter of interest in a finite set, and $\theta \in \Theta(a)$ continuous nuisance parameter in an open set. We also have iid observations that take values in $U$, a finite set. Let $p_{(a,\theta)}$ be a probability distribution on $U$. By the Bayesian methodology, we need to set up prior distributions for the parameters. Let $\pi(a)$ be the prior for $a$ which is discrete on $A$. For each $a \in A$, let $f_a(\theta)$ be the prior for $\theta$ on $\Theta(a)$. So, the Bayesian inference goes as follows: let $u = (u_1, ..., u_k) \in U^k$ iid data generated by $(a, \theta)$. Define the Maximum a Posteriori (MAP) by $\hat{a} = \arg\max_{b \in A} \pi(b) E_{\theta'}[P(u|b, \theta')]$ where the function to maximize corresponds to the posterior probability of $b|u$, $E_{\theta'}$ is the expected value taken with respect to $f_b(\theta)$ and $P(u|b, \theta') = \Pi_{j=1}^{k} p_{(b,\theta')}(u_i)$ is the likelihood of the data.

## 1.2 Phylogenetic trees notation

Translating the previous notation into the concepts seen in class, we will consider $A$ the set of fully resolved binary phylogenetic trees on a given leaf set. Let $U$ be set of possible site patterns and for each tree $a \in A$, $\Theta(a)$ corresponds to the branch lengths. So, for $n$ leaves, $\Theta(a) = (0, \infty)^{2n-3}$. Regarding the prior distribution, the usual choices are the uniform or Yule distribution for $\pi(a)$ and the exponential distribution for $f_a(\theta)$.

## 2 Main theorem

Provided the following conditions hold for each $a \in A$:
C1) $\pi(a) > 0$
C2) $f_a(\theta)$ is continuous bounded nonzero on $\Theta(a)$
C3) $\theta \longmapsto p_{(a,\theta)}$ is continuous nonzero on $\Theta(a)$
C4) $\forall \theta \in \Theta(a), b \neq a, \inf_{\theta' \in \Theta(b)} d(p_{(a,\theta)}, p_{(b,\theta')}) > 0$ where $d$ stands for the L1 metric.
Then, $\lim_{k\to\infty} P(a,\theta,k) = 1$ for all $a \in A$, $\theta \in \Theta(a)$ where $P(a,\theta,k)$ is the probability that the MAP estimator correctly selects $a$. Note that the priors in the phylogenetic notation satisfy (C1) and (C2). Also, (C3) is satisfied by any Markov process on a tree, and (C4) is satisfied by the identifiability of the model.

### Outline of the proof

The MAP estimator chooses $a$ from $u$ iff the Bayes factor is greater than 1 for all $b \neq a$. The Bayes factor is defined as $BF_{a/b} = \frac{\pi(a)E_\theta[P(u|a,\theta)]}{\pi(b)E_{\theta'}[P(u|b,\theta')]}$.

Note that $\pi(a)/\pi(b) > 0$ by (C1), so it suffices to prove that for all $b \neq a$ and $M < \infty$, $\lim_{k\to\infty} P(R_{a/b} > M) = 1$ for $R_{a/b} = \frac{E_\theta[P(u|a,\theta)]}{E_{\theta'}[P(u|b,\theta')]}$.

The idea of the proof is to find an explicit lower bound (LB) for the numerator of $R_{a/b}$ and an explicit upper bound (UB) for the denominator of $R_{a/b}$ such that $P(LB/UB > M)$ tends to 1 as $k \to \infty$.

It turns out that $R_{a/b} \geq \frac{\mu(N_\tau)\prod_{u\in U} s(u)^{r(u)k}}{\prod_{u\in U} q(u)^{r(u)k}}$ where $N_\tau$ is a closed ball of radius $\tau > 0$ centered at $\theta_0$ (the true parameter) that lies inside $\Theta(a)$, $\mu(N_\tau) = \int_{N_\tau} f_a(\theta)d\theta > 0$, $s(u)$ is the probability distribution of the form $p_{(a,\theta)}$ that minimizes $P(u|a,\theta)$ when $\theta$ is restricted to $N_\tau$, $r(u) = \frac{1}{k}n_u$ is the empirical probability distribution on $U$ with $n_u = |\{j : u_j = u\}|$, and $q(u)$ is the limit of the sequence $\{p_{(b,\theta_i)} : \theta_i \in \Theta(b), \lim_{i\to\infty} P(u|b,\theta_i) = \sup_{\theta'\in\Theta(b)} P(u|b,\theta')\}$.

The explanation of the bounds is skipped, but it comes from the fact that we can rewrite the likelihood as $P(u|b,\theta) = \prod_{u\in U} p_{(b,\theta)}(u)^{r(u)k}$. Taking logarithm to the bound, we get $log(R_{a/b}) \geq log[\mu(N_\tau)] + k\sum_{u\in U} r(u)log\frac{s(u)}{q(u)}$. So, it only remains to prove that $\sum_{u\in U} r(u)log\frac{s(u)}{q(u)} \geq \epsilon > 0$ regardless of $k, \tau$. For this purpose, the following lemma is needed whose conditions are satisfied by (C1)-(C4).

**Lemma:** $\forall \epsilon_1, \epsilon_2, \epsilon_3 > 0$, $\exists \delta, \epsilon > 0$ such that for $U$ finite set and probability distributions $p, q, r, s$ on $U$

i. $d(p,q) \geq \epsilon$

ii. $p(u) \geq \epsilon_2, q(u) \geq \epsilon_3, \forall u \in U$

iii. $d(p, r) < \delta, d(p, s) < \delta$

Then, $\sum_{u \in U} r(u) log \frac{s(u)}{q(u)} \geq \epsilon$.

The proof of the lemma is omitted, only the conditions are shown to be satisfied. Let $p = p_{(a,\theta_0)}$ the true probability distribution.

i. $\epsilon_1 = \inf_{\theta' \in \Theta(b)} d(p, p_{(b,\theta')})$ is positive by (C4). Since $q$ is of the form of $p_{(b,\theta')}$ (or a limit of such distributions), then $d(p, q) \geq \epsilon_1 > 0$.

ii. $\epsilon_2 = min\{p(u) : u \in U\}$ is positive by (C3). For $\rho > 0$, let $E_\rho$ be th event that $u$ is such that $d(p, r) < \rho$. Claim: for $\rho = \frac{1}{2}\epsilon_2$, $u$ satisfies $E_\rho$, and then, $q(u) \geq \epsilon_3 > 0$.

iii. Note that $d(p, r) \leq \delta$ is the event $E_\delta$. By LLN, $P(E_\delta) \to 1$ as $k \to \infty$. By (C3), we can choose $\tau > 0$ small such that $d(p, s) < \delta$.

Thus, the conditions on the lemma are satisfied which completes the proof of the theorem.

## References

[1] Steel, M. 2010. Consistency of Bayesian inference of resolved phylogenetic trees. arXiv:1001.2864v1