Assessing the stability of a tree: an analytic approach

M. Mariadassou   A. Bar-Hen

Department MMIP
AgroParisTech

2 July 07 / Hazing

Outline

Outline

1. Introduction
   - Motivations
   - Notations
   - Toy Example
2. Variability of the inferred phylogeny
   - Fluctuations of the mean log-likelihood
   - Inversion events
3. Comparison with bootstrap and further work
   - Bootstrap
   - Further work

The stability issue in phylogeny

- Inferred topology depends on the method and data;
- Focus on ML methods and their statistical properties;
- Variability of the topology induced by data sampling;
- Inferred clade may be erroneous;
- How confident are we in a clade ?

Notations I

Data structure:

- Data matrix $\mathcal{X} = (X_{ij})$ of size $s \times n$;
- X_{ij} character j in species i;
- Here, character = nucleotide and $X_{ij} \in \mathcal{A} = \{A, C, G, T\}$;
- X_i i-th column of $\mathcal{X}$, vector of size s;
- X_i nucleotide pattern of site i, *e.g.* $(AAATTT)'$;

Phylogenetic model T structure:

- an evolution model: substitution model and associated parameters
- a tree topology: branching pattern and branch lengths

Notations II

- X_i i.i.d. with shared distribution Q;
- empirical distribution $Q_n = \sum_i \delta_{X_i}$;
- True and empirical mean log-likelihood of T:

$$\ell^T = \mathbb{E}_Q[\log \mathbb{P}(X;T)] = \sum_{x \in \mathcal{A}^s} Q(x) \log \mathbb{P}(x;T) \quad (1)$$

$$\ell_n^T = \mathbb{E}_{Q_n}[\log \mathbb{P}(X;T)] = \frac{1}{n} \sum_i \log \mathbb{P}(X_i;T) \quad (2)$$

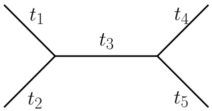
  where $\mathbb{P}(x;T)$ is the probability of pattern x under model T;
- True and empirical variance of T defined in the natural way.

Illustration with a toy example

Toy example made of $s = 4$ species and $n = 3$ nucleotides:

$$\mathcal{X} = \begin{array}{|c|ccc|} \hline \textit{Species} & & \textit{Sites} & \\ \hline S_1 & A & A & A \\ S_2 & G & G & C \\ S_3 & C & C & A \\ S_4 & C & C & C \\ \hline \end{array}$$

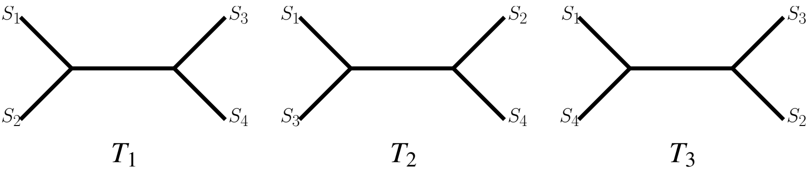
Topology with the following branch lenghts:



- First two nucleotides: identical and require at least a transition and a transversion
- Third nucleotide: minimal requirement is only a transversion.

Competing phylogenetic model

Three different topologies corresponding to the three possible labelings:

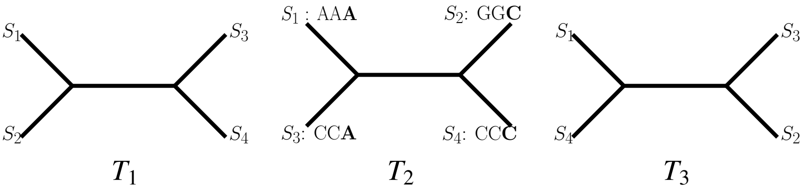


For a simple K2P model, with small α and β, we compare the three patterns in terms of likelihood and variance (among nucleotides).

Mean Likelihood and Variance of Nucleotides

For small values of α and β (K2P model),

- T_2: best tree in terms of mean likelihood, thanks to an outlier site;
- T_3: best tree in terms of variance, poor support from all sites;
- T_1: tree supported by 2 of the 3 nucleotides but selected by none of the above criteria;



ML based on the model ranking induced by their likelihood score (best tree = best likelihood score).

AgroParisTech

Outline

ℓ^T as a scalar product

- Replace Q and Q_n by $\boldsymbol{\theta} = (\theta^x)_{x \in \mathcal{A}^s}$ and $\boldsymbol{\theta}_n = (\theta_n^x)_{x \in \mathcal{A}^s}$:

$$\begin{aligned} \theta^x &= P_Q(X = x) \\ \theta_n^x &= P_{Q_n}(X = x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i = x\}} \end{aligned}$$

- Then, with $\log P^T = (\log P(x, T))_{x \in \mathcal{A}^s}$.

$$\begin{aligned} \ell^T &= \mathbb{E}_Q[\log P(X; T)] = \boldsymbol{\theta} . \log P^T \\ \ell_n^T &= \mathbb{E}_{Q_n}[\log P(X; T)] = \boldsymbol{\theta}_n . \log P^T \end{aligned}$$

- $\ell^T - \ell_n^T = (\boldsymbol{\theta} - \boldsymbol{\theta}_n) . \log P^T$

- To control $\ell^T - \ell_n^T$, we need to control $\theta - \theta_n$.

ℓ^T as a scalar product

- Replace Q and Q_n by $\boldsymbol{\theta} = (\theta^x)_{x \in \mathcal{A}^s}$ and $\boldsymbol{\theta}_n = (\theta_n^x)_{x \in \mathcal{A}^s}$:

$$\begin{aligned} \theta^x &= P_Q(X = x) \\ \theta_n^x &= P_{Q_n}(X = x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = x\}} \end{aligned}$$

- Then, with $\log P^T = (\log P(x, T))_{x \in \mathcal{A}^s}$.

$$\begin{aligned} \ell^T &= \mathbb{E}_Q[\log P(X; T)] = \boldsymbol{\theta} . \log P^T \\ \ell_n^T &= \mathbb{E}_{Q_n}[\log P(X; T)] = \boldsymbol{\theta}_n . \log P^T \end{aligned}$$

- $\ell^T - \ell_n^T = (\boldsymbol{\theta} - \boldsymbol{\theta}_n) . \log P^T$

- To control $\ell^T - \ell_n^T$, we need to control $\boldsymbol{\theta} - \boldsymbol{\theta}_n$.

Concentration Inequalities I

- By the law of large numbers, $\boldsymbol{\theta} - \boldsymbol{\theta}_n \longrightarrow_{n \to \infty} 0$;
- Probability of $\{|\boldsymbol{\theta} - \boldsymbol{\theta}_n| > \epsilon\}$ decreases exponentially towards 0;
- At what rate ?

Measure concentration tools give:

$$\frac{\log \mathbb{P}(\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| > \varepsilon)}{n} \leq \frac{s}{n} \log |\mathcal{A}| + \frac{\log 2}{n} + \max_{x \in \mathcal{A}^s} \frac{-\varepsilon^2}{\theta^x(1 - \theta^x + \varepsilon)} \quad (3)$$

Concentration Inequalities II

Using (3), we derive:

$$\frac{\log \mathbb{P}\left(|\ell^T - \ell_n^T| \geq \varepsilon\right)}{n} \leq \frac{s}{n}\log|\mathcal{A}| + \frac{\log 2}{n} + \max_{x\in\mathcal{A}^s} \frac{-\tilde{\varepsilon}^2}{\theta^x(1-\theta^x+\tilde{\varepsilon})}$$

Where $\tilde{\varepsilon} = \dfrac{\varepsilon}{|\mathcal{A}|^s \| \log P^T \|_\infty}$.

Remarks:

- The exponential rate of decay $\tilde{\varepsilon}$ is very low but....
- $\mathcal{A}^s$ is too big (impossible patterns)
- Better bounds when replacing $\mathcal{A}^s$ by the true number of possible patterns
- Sharp bound for extreme θ^x but not medium ones ($\simeq 1/2$).
- For a given confidence level, we know how n evolve with s.

Concentration Inequalities II

Using (3), we derive:

$$\frac{\log \mathbb{P}\left(|\ell^T - \ell_n^T| \geq \varepsilon\right)}{n} \leq \frac{s}{n}\log|\mathcal{A}| + \frac{\log 2}{n} + \max_{x\in\mathcal{A}^s} \frac{-\tilde{\varepsilon}^2}{\theta^x(1-\theta^x+\tilde{\varepsilon})}$$

Where $\tilde{\varepsilon} = \dfrac{\varepsilon}{|\mathcal{A}|^s \|\log P^T\|_\infty}$.

Remarks:

- The exponential rate of decay $\tilde{\varepsilon}$ is very low but....
- $\mathcal{A}^s$ is too big (impossible patterns)
- Better bounds when replacing $\mathcal{A}^s$ by the true number of possible patterns
- Sharp bound for extreme θ^x but not medium ones ($\simeq 1/2$).
- For a given confidence level, we know how n evolve with s.

Concentration Inequalities II

Using (3), we derive:

$$\frac{\log \mathbb{P}\left(|\ell^T - \ell_n^T| \geq \varepsilon\right)}{n} \leq \frac{s}{n} \log |\mathcal{A}| + \frac{\log 2}{n} + \max_{x \in \mathcal{A}^s} \frac{-\tilde{\varepsilon}^2}{\theta^x (1 - \theta^x + \tilde{\varepsilon})}$$

Where $\tilde{\varepsilon} = \dfrac{\varepsilon}{|\mathcal{A}|^s \| \log P^T \|_\infty}$.

Remarks:

- The exponential rate of decay $\tilde{\varepsilon}$ is very low but....
- $\mathcal{A}^s$ is too big (impossible patterns)
- Better bounds when replacing $\mathcal{A}^s$ by the true number of possible patterns
- Sharp bound for extreme θ^x but not medium ones ($\simeq 1/2$).
- For a given confidence level, we know how n evolve with s.

Inversions events

- ML methods based on the model ranking induced by their likelihood score;
- But inference done on ranking induced by empirical likelihood score;
- Inversion events can appear:
- When comparing two models T and T', the true ranking may be different from the empirical one;
- How often does such an event happens ?
- How does its probability decreases when available information increases ?

Inversions events

- ML methods based on the model ranking induced by their likelihood score;
- But inference done on ranking induced by empirical likelihood score;
- Inversion events can appear:
- When comparing two models T and T', the true ranking may be different from the empirical one;
- How often does such an event happens ?
- How does its probability decreases when available information increases ?

Concentration results

Using the same concentration tools, we obtain:

Proposition

Assume that model T is better than model T' ($\ell^T > \ell^{T'}$), then the probability that T' is better than T for our sample is such that:

$$\frac{\log \mathbb{P}(\ell_n^T - \ell_n^{T'} < 0)}{n} \leq \frac{s}{n} \log |\mathcal{A}| + \max_{x \in \mathcal{A}^s} \frac{-\varepsilon^2}{\theta^x (1 - \theta^x + \varepsilon)} \qquad (4)$$

where $\varepsilon = \frac{\ell^T - \ell^{T'}}{|\mathcal{A}^s| \| \log P^T - \log P^{T'} \|}$

- The same remark as before apply, the bound is considerably sharper when replacing $|\mathcal{A}|^s$ by the true number of patterns.
- Comforting result: inversion probability decreases with $\ell^T - \ell^{T'}$.

AgroParisTech

Outline

1. Introduction
   - Motivations
   - Notations
   - Toy Example
2. Variability of the inferred phylogeny
   - Fluctuations of the mean log-likelihood
   - Inversion events
3. Comparison with bootstrap and further work
   - Bootstrap
   - Further work

Limitations of Bootstrap

- Same concept as the bootstrap: evaluate the probability of the inferred true not being the true one;
- Bootstrap relies heavily on simulations;
- The only variability is the observed variability (non-parametric bootstrap);
- Significance threshold decided *ex-ante* (66% or 95%);
- No justification for the threshold;
- No link between n and s.

Advantages of the analytical bound

- Focus on the likelihood score (instead of the phylogeny and/or topology);
- Accounts for more variability than just the one observed in the data;
- Accounts for s and n when calculating a confidence level;
- For a given s and a given confidence level, calculate the necessary number of sites;
- No heavy computations.

Perspectives

- Develop a plug-in estimator of the confidence level of a phylogenetic model;
- Compare our method with boostrap on data (real and/or simulated, toy example);
- Consider process bounds instead of pointwise bounds;
- Anything else I can think about.