ERaBLE: branch lengths estimation for phylogenomic trees

June 7, 2017

Arthur Zwaenepoel
Master of science in Bioinformatics: Systems Biology

Introduction

One of the central problems in phylogenomics is the inference of a species phylogeny based on a large set of homologous genomic information. Several approaches have been developed to infer a species tree from a set of genes tree, but virtually none of them are able to provide meaningful branch lengths, and are therefore restricted to topology inference only. Binet *et al.* (2016) [1] proposed a method for estimating branch lengths of a species tree based on gene-wise distance matrices. Besides branch lengths, also gene-wise evolutionary rates are determined. The method is based on the solution of an equality constrained convex optimization problem, for which they described and implemented an analytical solution. Their method was called ERaBLE, for evolutionary rates and branch lengths estimation.

Optimization problem

Given a species tree $\mathcal{T}$ and a set of genes $G^{(1)}, G^{(2)} \cdots, G^{(m)}$, let $\delta_{ij}^{(k)}$ be the distance between sequence $G^{(k)}$ in taxon i and taxon j. The topology matrix A is defined as the $n(n-1)/2 \times \tau$ matrix with pairs of taxa in the rows and all branches of the tree $\mathcal{T}$ in the columns, an entry in the matrix is set to one if the corresponding branch is part of the path between the two taxa and to zero otherwise. Let $\hat{d}$ be the vector of additive distances between pairs of taxa, obtained by multiplying the topology matrix A with the vector of estimated branch lengths $\hat{b}$. Let $\hat{\alpha}_k$ be the scale factor defined as $1/r_k$, with r_k the evolutionary rate for G_k. Finally let L_k be the subset of taxa for which G_k is present. The optimization problem solved in ERaBLE can be expressed as:

$$\begin{aligned} \underset{\hat{\alpha},\hat{b}}{\text{minimize}} \quad & \sum_{k=1}^{m} \sum_{\{i,j\}\subset L_k} w_{ij}^{(k)} (\hat{\alpha}_k \delta_{ij} - \hat{d}_{ij})^2 \\ \text{subject to} \quad & \sum_{k=1}^{m} Z_k \hat{\alpha}_k = \sum_{k=1}^{m} Z_k \end{aligned} \tag{1}$$

Which is a weighted least squares (WLS) like problem with weights w_{ij}. The difference with normal WLS is the estimation of the scale factors α_k, which is required since the gene-wise distances are not directly comparable to the species distances. The rationale behind the objective function is that we would like to set the additive distances in the species tree equal to those in in the gene tree up to a proportionality constant ($\hat{\alpha}_k$). This is the so-called *proportional* model, an assumption that is almost certainly violated due to heterotachy and other issues. The equality constraint is included because the objective function is trivially minimized by setting $\hat{\alpha}$ and $\hat{b}$ to zero (recall $\hat{d} = A\hat{b}$). The constraint prevents $\hat{\alpha}_k = 0 \ \forall k \in \{1, ..., m\}$ as a solution. The authors chose Z_k as $N_k \sum_{i,j \in |L_k|} \delta_{ij}^{(k)}$ and $w_{ij}^{(k)} = N_k$.

Solution

The problem is a quadratic programming problem and has an analytical solution which can be found using the method of Lagrange multipliers. The Lagrangian function is defined as

$$\mathcal{L}(\hat{\alpha}, \hat{b}, \lambda) = Q(\hat{\alpha}, \hat{b}) + \lambda(z^T \hat{\alpha} - Z) \tag{2}$$

with Q the objective function, z^T the vector $(Z_1, Z_2, \cdots, Z_m)^T$ and $Z = \sum_{i=1}^{m} Z_m$. As a sum of squares is convex and the constraint is linear, a necessary and sufficient condition for optimality is given by $\nabla_{\hat{\alpha},\hat{b},\lambda} \mathcal{L}(\hat{\alpha}, \hat{b}, \lambda) = 0$. If we rewrite the objective function as:

$$Q(\hat{\alpha}, \hat{b}) = \sum_{k=1}^{m} (\hat{\alpha}_k \delta_k - A_k \hat{b})^T W_k (\hat{\alpha}_k \delta_k - A_k \hat{b}) \tag{3}$$

The solution can be derived as the following system

$$\begin{cases} \hat{\alpha}_k \delta_k^T W_k \delta_k - \delta^T W_k A_k b + \lambda z^T/2 &= 0 \\ \sum\limits_{k=1}^{m} (A_k^T W_k A_k \hat{b} - \hat{\alpha}_k A_k^t W_k \delta_k) &= 0 \\ z^T \hat{\alpha} - Z &= 0 \end{cases} \tag{4}$$

Which is a system of $m+\tau+1$ equations, where m is the number of genes, and τ the number branches in the tree. This system can be written as

$$\begin{cases} D\hat{\alpha} + B^T \hat{b} + \lambda z &= 0 \\ B\hat{\alpha} + C\hat{b} &= 0 \\ z^T \hat{\alpha} &= Z \end{cases} \tag{5}$$

Which can be solved naively in $O(mn^4 + (n+m)^3)$ time, but cleverly in $O(mn^2 + n^3)$, with n the number of taxa. Because of the proportionality assumption, where we assume that the true gene-wise rates have a weighted average of 1, both the estimated rates $\hat{\alpha}$ as well as the phylogenomic branch lengths are subsequently rescaled using the scaling factor $c = 1/N \sum_{k=1}^{m} N_k/\hat{\alpha}_k$. This ensures that the actual size of the constraint Z does not influence the final results.

Python implementation

A Python implementation in the form of a Jupyter notebook can be retrieved from `https://github.com/arzwa/erable`.

References

[1] Manuel Binet, Olivier Gascuel, Celine Scornavacca, Emmanuel J. P. Douzery, and Fabio Pardi. Fast and accurate branch lengths estimation for phylogenomic trees. *BMC Bioinformatics*, 17(1):23, 2016.