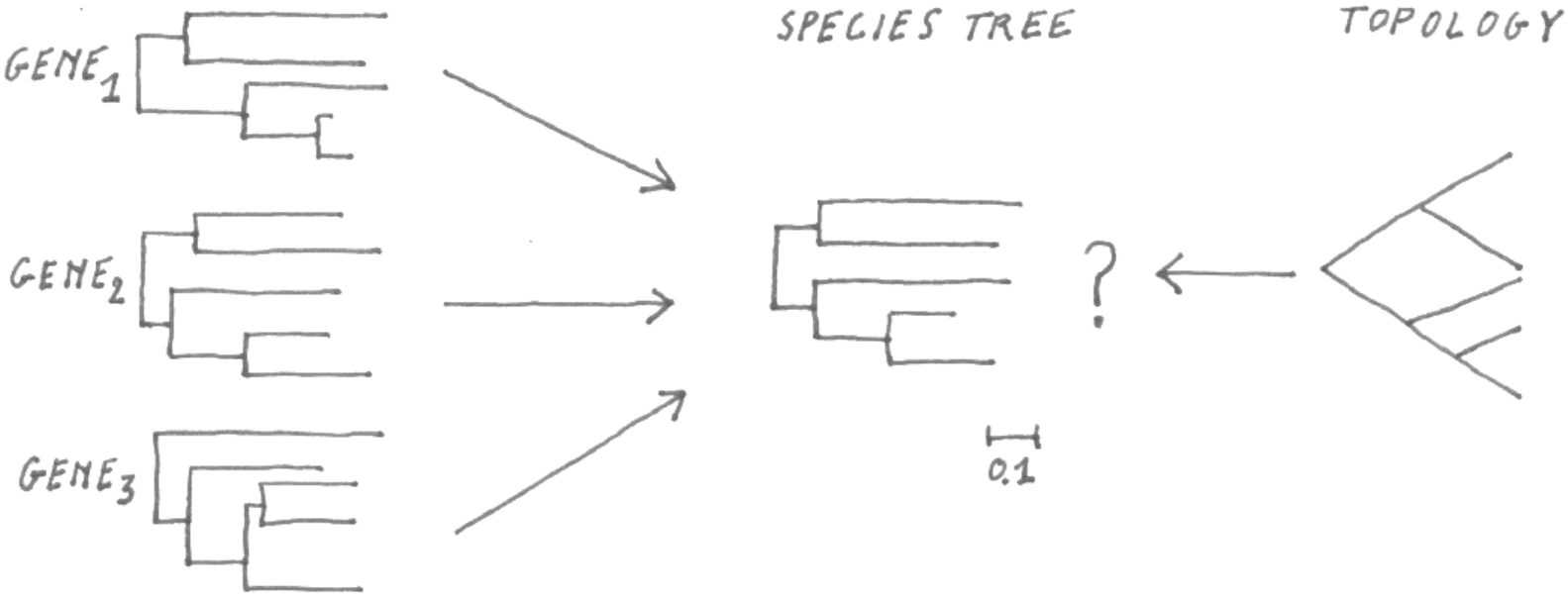
ERaBLE: branch lengths estimation for phylogenomic trees

Arthur Zwaenepoel

Msc. Bioinformatics: Systems Biology, Ghent University

June 8, 2016

Setting: phylogenomics



$$b_e = \frac{\sum_{k=1}^{m} s_{xy}^{(k)}}{\sum_{k=1}^{m} N_k} \qquad b_e = \frac{1}{N} \sum_{k=1}^{m} N_k \, b_e^{(k)}$$

Approaches

- Supertree
- Superalignment
- Medium-level

Assumptions

- Gene trees are topologically compatible, variation due to rate heterogeneity
- Proportional model

$$\frac{b_e^{(k)}}{r_k} = \text{constant} = b_e \qquad \text{for all } k = 1, 2, \cdots, m$$

Problem

Estimate gene-wise evolutionary rates $\hat{\alpha}_k$ and branch lengths $\hat{b}$ for a given species tree topology $\mathcal{T}$ and a set of distances matrices δ for m genes $(G_1, G_2, \cdots, G_m)$

$$\hat{d}_{ij} \approx \hat{\alpha}_k \delta_{ij}^{(k)}$$

Where $\hat{d} = A_k \hat{b}$, with A the topology matrix for tree $\mathcal{T}$, *i.e.* the additive distance between taxon i and j.

Sidenote: topology matrix

The topology matrix A is the $\frac{|\mathcal{T}|(|\mathcal{T}|-1)}{2} \times \tau$ binary matrix

$$A = \begin{array}{c} \\ \textit{species}1-2 \\ \textit{species}1-3 \\ \vdots \\ \textit{species}2-4 \\ \textit{species}3-4 \end{array} \begin{array}{c} \begin{array}{cccc} b_1 & b_2 & \cdots & b_6 \end{array} \\ \begin{pmatrix} 1 & 1 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \end{pmatrix} \end{array}$$

A_k is as A but with rows of species that lack G_k omitted

Problem

Estimate gene-wise evolutionary rates $\hat{\alpha}_k$ and branch lengths $\hat{b}$ for a given species tree topology $\mathcal{T}$ and a set of distances matrices δ for m genes $(G_1, G_2, \cdots, G_m)$

$$\begin{aligned} \underset{\hat{\alpha},\hat{b}}{\text{minimize}} \quad & \sum_{k=1}^{m} (\hat{\alpha}_k \delta_k - A_k \hat{b})^T W_k (\hat{\alpha}_k \delta_k - A_k \hat{b}) \\ \text{subject to} \quad & \sum_{k=1}^{m} Z_k \hat{\alpha}_k = \sum_{k=1}^{m} Z_k \end{aligned} \tag{1}$$

Where $A_k \hat{b} = \hat{d}$, with A_k the topology matrix for tree $\mathcal{T}$ over taxa that have G_k.

Solution

Lagrangian

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

Sum of squares is convex and constraint is linear so a necessary and sufficient condition for optimality is given by

$$\nabla_{\hat{\alpha}, \hat{b}, \lambda} \mathcal{L}(\hat{\alpha}, \hat{b}, \lambda) = 0$$

Which can be solved naively in $O(mn^4 + (n+m)^3)$ and can be further brought down to $O(mn^2 + n^3)$

Computational efficiency

Table 2 Computational efficiencies on the OrthoMaM data set for the tested methods

	Concat+Dist	Concat+ML	SDM*add	DistRadd	ERaBLEadd	SDM*	DistR	ERaBLE
T_1	≈ 0		3 h 20 m/39 h 28 m			2 m 46 s	2 m 46 s	2 m 46 s
T_2	5 m 41 s	41 h 16 m	8 h 2 m	2 h 9 m	7 s	8 h 33 m	2 h 6 m	7 s
M	889 MB	117 GB	1.2 GB	2.8 GB	222 MB	1.2 GB	3.0 GB	221 MB

Implementation & Reference

Python implementation (jupyter notebook):

- `https://github.com/arzwa/erable`

Reference

- Binet, M., Gascuel, O., Scornavacca, C., P. Douzery, E. J., & Pardi, F. (2016). Fast and accurate branch lengths estimation for phylogenomic trees. BMC Bioinformatics, 17(1), 23. https://doi.org/10.1186/s12859-015-0821-8