Fast and accurate branch length estimation for phylogenomic trees: ERaBLE (Evolutionary Rates and Branch Length Estimation)

Manuel Binet, Olivier Gascuel, Celine Scornavacca, Emmanuel J.P. Douzery, Fabio Pardi

Rencontres Alphy
Génomique Evolutive, Bioinformatique, Alignement et Phylogénie

Fast and accurate branch length estimation for phylogenomic trees: ERaBLE

(Evolutionary Rates and Branch Length Estimation).

Manuel Binet, Olivier Gascuel, Celine Scornavacca, Emmanuel J.P. Douzery and Fabio Pardi

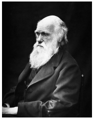
March 10, 2015

Laboratoire Informatique Robotique Microélectronique Montpellier

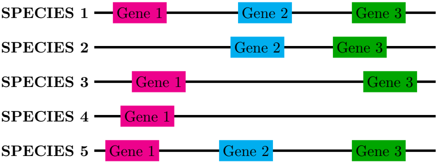
IBC

isem
Institut des Sciences de l'Evolution-Montpellier

The Phylogenomics era



Important goal: build the species tree.



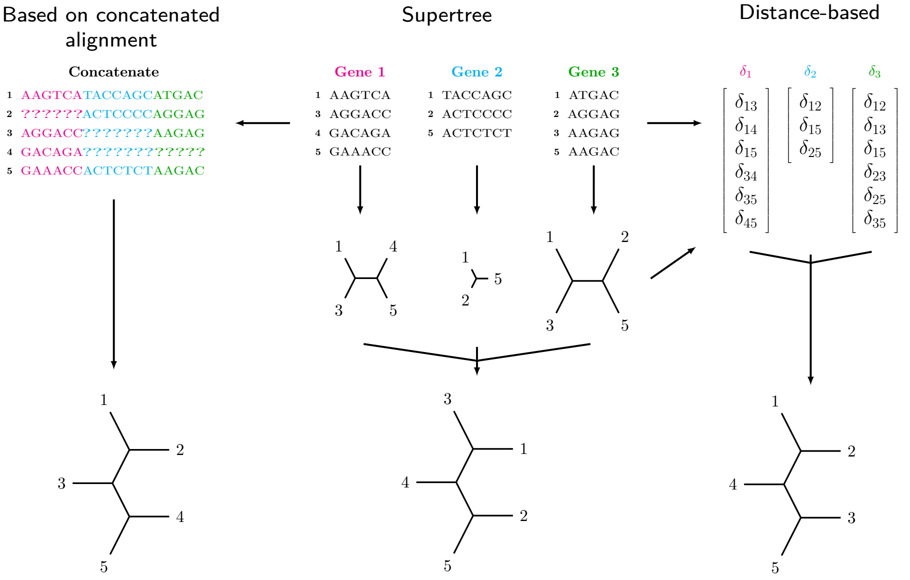
New challenges:

- Use the maximum of available data.
- Account for gene rate heterogeneity.
- Deal with conflicts in gene trees topologies.

  ...

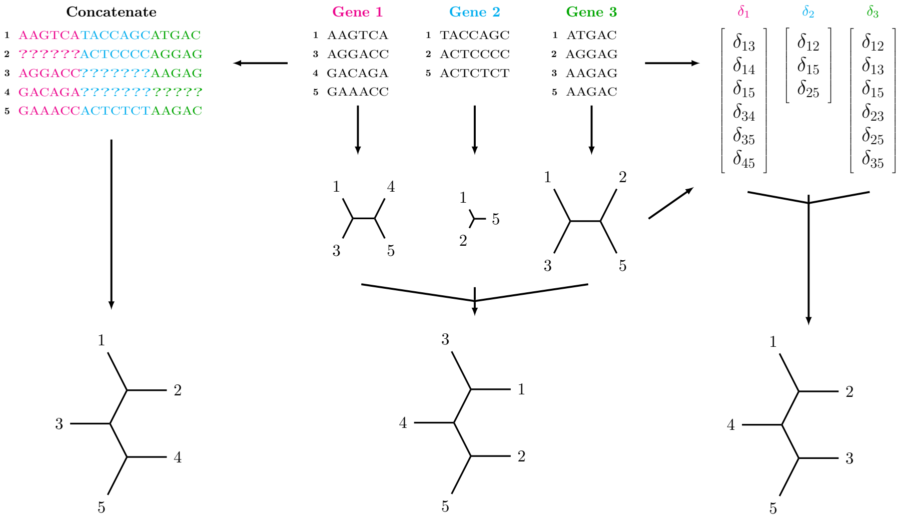
ERaBLE (Evolutionary Rates and Branch Lengths estimation).

Categories of phylogenomics methods



Categories of phylogenomics methods

Based on concatenated alignment | Supertree | Distance-based



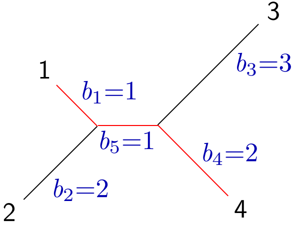
Branch lengths:

yes but computationally heavy | usually no | yes

Tree distances

Every tree with branch lengths can be represented with a vector of tree distances by computing the distance d_{ij}^T between each pair of taxa in the tree.

$T=(\mathcal{T}, b)$



$\Leftrightarrow$

$$d^T = \begin{pmatrix} d_{12} = 3 \\ d_{13} = 5 \\ \mathbf{d_{14} = 4} \\ d_{23} = 6 \\ d_{24} = 5 \\ d_{34} = 5 \end{pmatrix}$$

Vector of tree distances

$$d_{ij}^T = \sum_{e \in P_{ij}} b_e$$

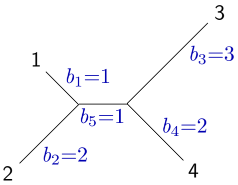
Tree distances

Every vector of tree distances is the product of the topological matrix of T by the branch lengths vector of T. i.e. $d^T = Ab$.

$$d^T = \begin{pmatrix} d_{12} = 3 \\ d_{13} = 5 \\ d_{14} = 4 \\ d_{23} = 6 \\ d_{24} = 5 \\ d_{34} = 5 \end{pmatrix} \qquad A = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix} \qquad b = \begin{pmatrix} b_1 = 1 \\ b_2 = 2 \\ b_3 = 3 \\ b_4 = 2 \\ b_5 = 1 \end{pmatrix}$$

Vector of tree distances — Topological matrix — Branch lengths vector

T=$(\mathcal{T}, b)$



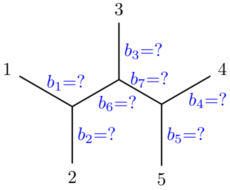
$$d^T = Ab$$

A distance-based method : the weighted least square method (WLS)

Branch lengths estimation with WLS:

$$\delta = \begin{bmatrix} \delta_{12} = 0.140 \\ \delta_{13} = 0.163 \\ \delta_{14} = 0.288 \\ \delta_{15} = 0.336 \\ \delta_{23} = 0.188 \\ \delta_{24} = 0.413 \\ \delta_{25} = 0.411 \\ \delta_{34} = 0.298 \\ \delta_{35} = 0.213 \end{bmatrix}$$

input distances
(estimated from
1 gene alignment)



input topology

$$d^T = Ab$$

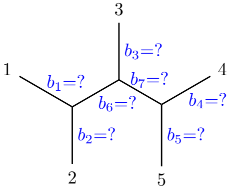
$$\text{minimize} \sum_{ij} w_{ij}(\delta_{ij} - d_{ij}^T)^2$$

A distance-based method : the weighted least square method (WLS)

Branch lengths estimation with WLS:

$$\delta = \begin{bmatrix} \delta_{12} = 0.140 \\ \delta_{13} = 0.163 \\ \delta_{14} = 0.288 \\ \delta_{15} = 0.336 \\ \delta_{23} = 0.188 \\ \delta_{24} = 0.413 \\ \delta_{25} = 0.411 \\ \delta_{34} = 0.298 \\ \delta_{35} = 0.213 \end{bmatrix}$$

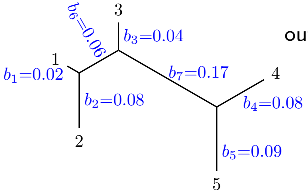
input distances
(estimated from
1 gene alignment)



input topology

$$d^T = Ab$$

minimize $\sum_{ij} w_{ij}(\delta_{ij} - d_{ij}^T)^2$

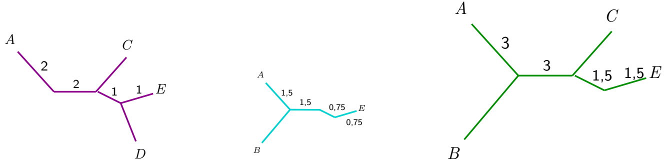


output

can be solved analytically in $\mathcal{O}(n^3)$

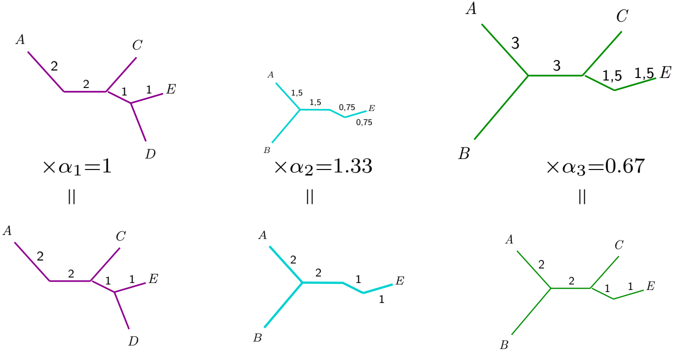
New methode ERaBLE: Evolutionnary Rates and Branch Length Estimation

Hypothesis: Any gene G_k induces approximately the same tree up to a scale factor α_k and the removal of a number of lineages.



New methode ERaBLE: Evolutionnary Rates and Branch Length Estimation

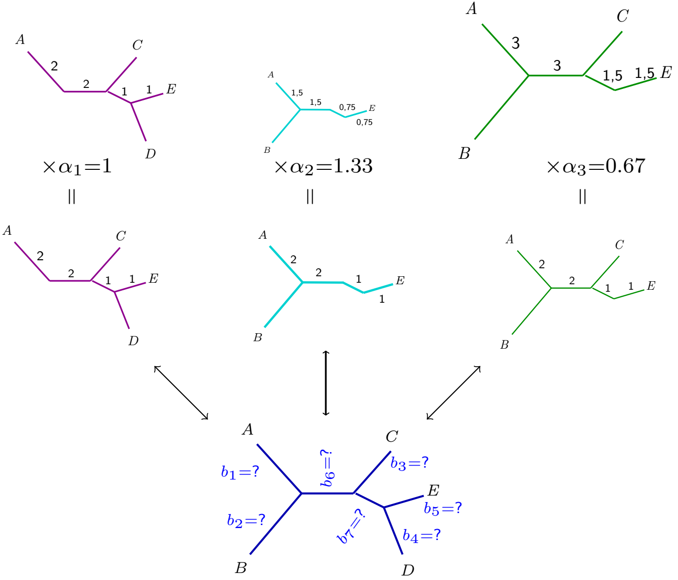
Hypothesis: Any gene G_k induces approximately the same tree up to a scale factor α_k and the removal of a number of lineages.



After scaling

New methode ERaBLE: Evolutionnary Rates and Branch Length Estimation

Hypothesis: Any gene G_k induces approximately the same tree up to a scale factor α_k and the removal of a number of lineages.



After scaling

New methode ERaBLE: Evolutionnary Rates and Branch Length Estimation

Input data: m distance vectors $\delta_1, \ldots, \delta_m$ where $\delta_k = (\delta_{ij}^{(k)})$ is defined on the taxa set L_k.

A given topology $\mathcal{T}$ defined on the taxa set $L = \displaystyle\bigcup_{k=1}^{m} L_k$.

New methode ERaBLE: Evolutionnary Rates and Branch Length Estimation

Input data: m distance vectors $\delta_1, \ldots, \delta_m$ where $\delta_k = (\delta_{ij}^{(k)})$ is defined on the taxa set L_k.

A given topology $\mathcal{T}$ defined on the taxa set $L = \bigcup_{k=1}^{m} L_k$.

Objective: find

- the branch lengths $\hat{b}_1, \ldots, \hat{b}_{2n-3}$ of $\mathcal{T}$,
- the scale factors $\hat{\alpha}_1, \ldots, \hat{\alpha}_m$ of the m genes,

solution of the problem:

$$\left\{ \begin{array}{ll} \text{minimize} & \displaystyle\sum_{k=1}^{m} \sum_{i,j \in L_k} w_{ij}^{(k)} (\hat{\alpha}_k \delta_{ij}^{(k)} - d_{ij}^{T})^2 \\ & \\ & \end{array} \right.$$

New methode ERaBLE: Evolutionnary Rates and Branch Length Estimation

Input data: m distance vectors $\delta_1, \ldots, \delta_m$ where $\delta_k = (\delta_{ij}^{(k)})$ is defined on the taxa set L_k.

A given topology $\mathcal{T}$ defined on the taxa set $L = \bigcup_{k=1}^{m} L_k$.

Objective: find

- the branch lengths $\hat{b}_1, \ldots, \hat{b}_{2n-3}$ of $\mathcal{T}$,
- the scale factors $\hat{\alpha}_1, \ldots, \hat{\alpha}_m$ of the m genes,

solution of the problem:

$$\begin{cases} \text{minimize} & \displaystyle\sum_{k=1}^{m} \sum_{i,j \in L_k} w_{ij}^{(k)} (\hat{\alpha}_k \delta_{ij}^{(k)} - d_{ij}^{T})^2 \\ \text{subjet to} & \displaystyle\sum_{k=1}^{m} Z_k \hat{\alpha}_k = 1 \end{cases}$$

New methode ERaBLE: Evolutionnary Rates and Branch Length Estimation

Input data: m distance vectors $\delta_1, \ldots, \delta_m$ where $\delta_k = (\delta_{ij}^{(k)})$ is defined on the taxa set L_k.

A given topology $\mathcal{T}$ defined on the taxa set $L = \bigcup_{k=1}^{m} L_k$.

Objective: find

- the branch lengths $\hat{b}_1, \ldots, \hat{b}_{2n-3}$ of $\mathcal{T}$,
- the scale factors $\hat{\alpha}_1, \ldots, \hat{\alpha}_m$ of the m genes,

solution of the problem:

$$\begin{cases} \text{minimize} & \displaystyle\sum_{k=1}^{m} \sum_{i,j \in L_k} w_{ij}^{(k)} (\hat{\alpha}_k \delta_{ij}^{(k)} - d_{ij}^{T})^2 \\ \text{subjet to} & \displaystyle\sum_{k=1}^{m} Z_k \hat{\alpha}_k = 1 \end{cases}$$

Outputs:

$$\hat{b}_e \quad \text{et} \quad \hat{r}_k = \frac{1}{\hat{\alpha}_k}$$

.

Resolution of the optimization problem

The minimization problem can be solved with the method of Lagrange multipliers and leads to the linear system in $\mathcal{O}(n+m)$ equations and unknowns:

$$\begin{pmatrix} \delta_1^t W_1 \delta_1 & 0 & \cdots & 0 & [\ -\delta_1^t W_1 A_1\ ] & 1 \\ 0 & \delta_2^t W_2 \delta_2 & & \vdots & [\ -\delta_2^t W_2 A_2\ ] & 1 \\ \vdots & & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \delta_m^t W_m \delta_m & [-\delta_m^t W_m A_m\ ] & 1 \\ [-A_1^t W_1 \delta_1] & [-A_2^t W_2 \delta_2] & \cdots & [-A_m^t W_m \delta_m] & \left[\sum_{k=1}^{m} A_k^t W_k A_k \right] & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ Z_1 & Z_2 & \cdots & Z_m & 0 \ \cdots \ 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \\ b_1 \\ b_2 \\ \vdots \\ b_{2n-3} \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

- Filling the matrix and standard system solving in $\mathcal{O}((n+m)^3 + mn^4)$
- Solving in $\mathcal{O}(n^3 + mn^2)$ with our algorithms:
  - Filling the matrix in $\mathcal{O}(mn^2)$ instead of $\mathcal{O}(mn^4)$
  - Block-solving system in $\mathcal{O}(n^3)$ instead of $\mathcal{O}(n+m)^3$

m=# genes
n=# taxa

Resolution of the optimization problem

The minimization problem can be solved with the method of Lagrange multipliers and leads to the linear system in $\mathcal{O}(n+m)$ equations and unknowns:

$$
\begin{pmatrix}
\delta_1^t W_1 \delta_1 & 0 & \cdots & 0 & [\ -\delta_1^t W_1 A_1\ ] & 1 \\
0 & \delta_2^t W_2 \delta_2 & & \vdots & [\ -\delta_2^t W_2 A_2\ ] & 1 \\
\vdots & & \ddots & \vdots & \vdots & \vdots \\
0 & \cdots & \cdots & \delta_m^t W_m \delta_m & [-\delta_m^t W_m A_m\ ] & 1 \\
\left[-A_1^t W_1 \delta_1 \right] & \left[-A_2^t W_2 \delta_2 \right] & \cdots & \left[-A_m^t W_m \delta_m \right] & \left[\sum_{k=1}^{m} A_k^t W_k A_k \right] & \begin{matrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{matrix} \\
Z_1 & Z_2 & \cdots & Z_m & 0 \ \cdots \ 0 & 0
\end{pmatrix}
\begin{pmatrix}
\alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \\ b_1 \\ b_2 \\ \vdots \\ \vdots \\ \vdots \\ b_{2n-3} \\ \lambda
\end{pmatrix}
=
\begin{pmatrix}
0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 0 \\ 1
\end{pmatrix}
$$

- Filling the matrix and standard system solving in $\mathcal{O}((n+m)^3 + mn^4)$
- Solving in $\mathcal{O}(n^3 + mn^2)$ with our algorithms:
  - Filling the matrix in $\mathcal{O}(mn^2)$ instead of $\mathcal{O}(mn^4)$
  - Block-solving system in $\mathcal{O}(n^3)$ instead of $\mathcal{O}(n+m)^3$

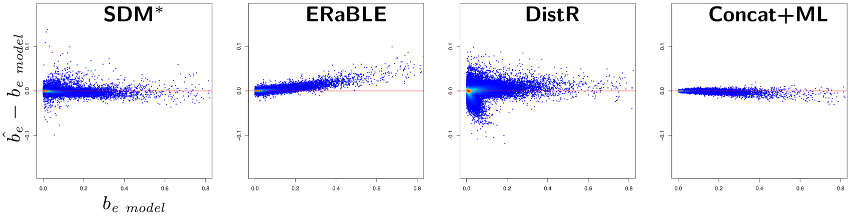
m=# genes
n=# taxa

Complexity: ERaBLE $\mathcal{O}(n^3 + mn^2)$ vs. WLS $\mathcal{O}(n^3)$

Results on a simulated dataset

input data: m=500 genes defined on $n \in [4, 40]$ taxa + model topology (500 replicates).

Accuracy in the estimation of branch lengths:



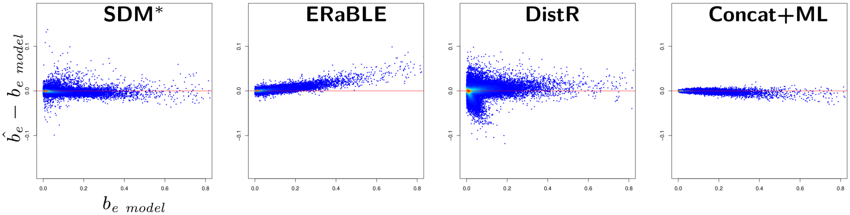
Compared methods (after adaptation):

- SDM* [Criscuolo et al. 2006]. Phylogenomic distance-based method,
- DistR [Bevan et al. 2005]. Distance-based method for the estimation of gene rates,
- Concat+ML. PhyML [Guindon et al. 2010] analysis of the concatenated alignments.

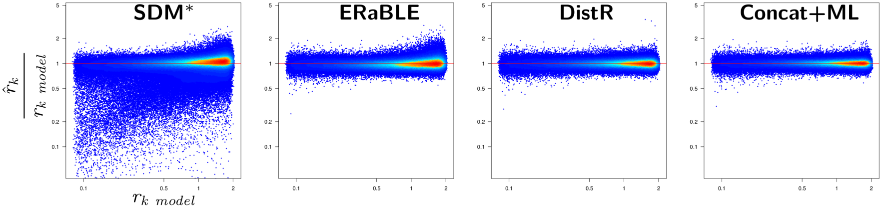
Results on a simulated dataset

input data: m=500 genes defined on $n \in [4, 40]$ taxa + model topology (500 replicates).

Accuracy in the estimation of branch lengths:



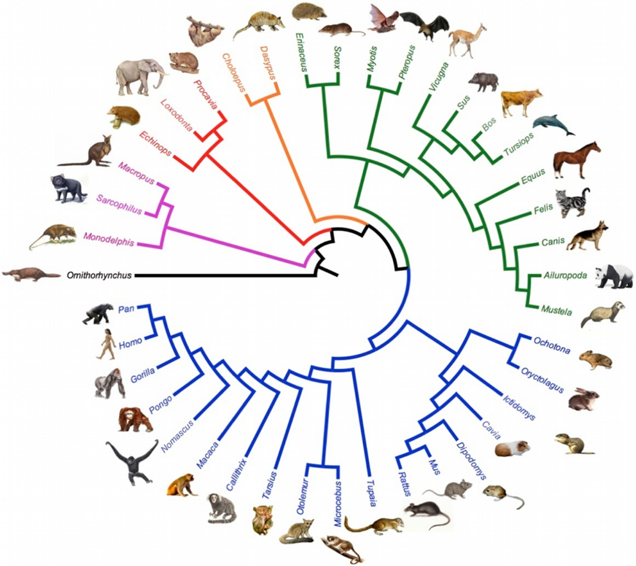
Accuracy in the estimation of gene rates (logarithmic scale):



Results on the OrthoMaM dataset [Douzery et al. 2014]

input data: m=6953 nucleotide exon alignments over n=4 to 40 mammals.

input topology: the topology of the 40 mammals present in OrthoMaM:

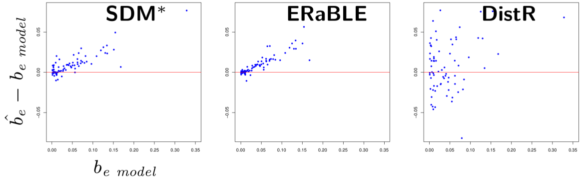


Results on the OrthoMaM dataset

	SDM*	ERaBLE	DistR	Concat+ML
Running times	8h33m⋆	7s⋆	2h8m⋆	41h16m
RAM	1.2 GB	221 MB	3 GB	488 GB

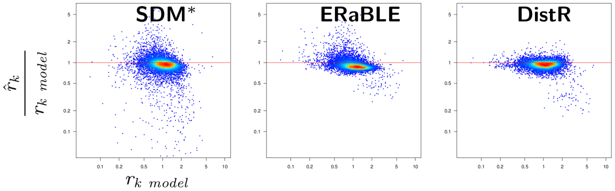
⋆ add 2min46s for the 6953 input distances estimation

Accuracy in the estimation of branch lengths:



Concat+ML as model

Accuracy in the estimation of gene rates (logarithmic scale):



Concat+ML as model

Conclusion

ERaBLE gives branch lengths to phylogenomic trees (e.g. estimated with supertree methods).

ERaBLE is relatively accurate in the estimations of branch lengths and of gene rates.

ERaBLE is fast with a complexity in $\mathcal{O}(n^3 + mn^2)$ (linear in m).

Thank you for your attention.

Any questions?

Fundings: