

Reconstructing the evolutionary history of a B cell lineage with minimum spanning tree and genotype abundances Supplementary File

Nika Abdollahi¹, Lucile Jeusset^{1,2}, Anne de Septenville², Frédéric Davi^{2,*} and Juliana Silva Bernardes^{1,*}

¹Sorbonne Université, CNRS, UMR 7238, Laboratoire de Biologie Computationnelle et Quantitative, Paris, France

²Sorbonne Université, AP-HP, Hôpital Pitié-Salpêtrière, Department of Biological Hematology, Paris, France

*To whom correspondence should be addressed; E-mail: juliana.silva_bernardes@upmc.fr.

October 10, 2022

Supplementary Figures

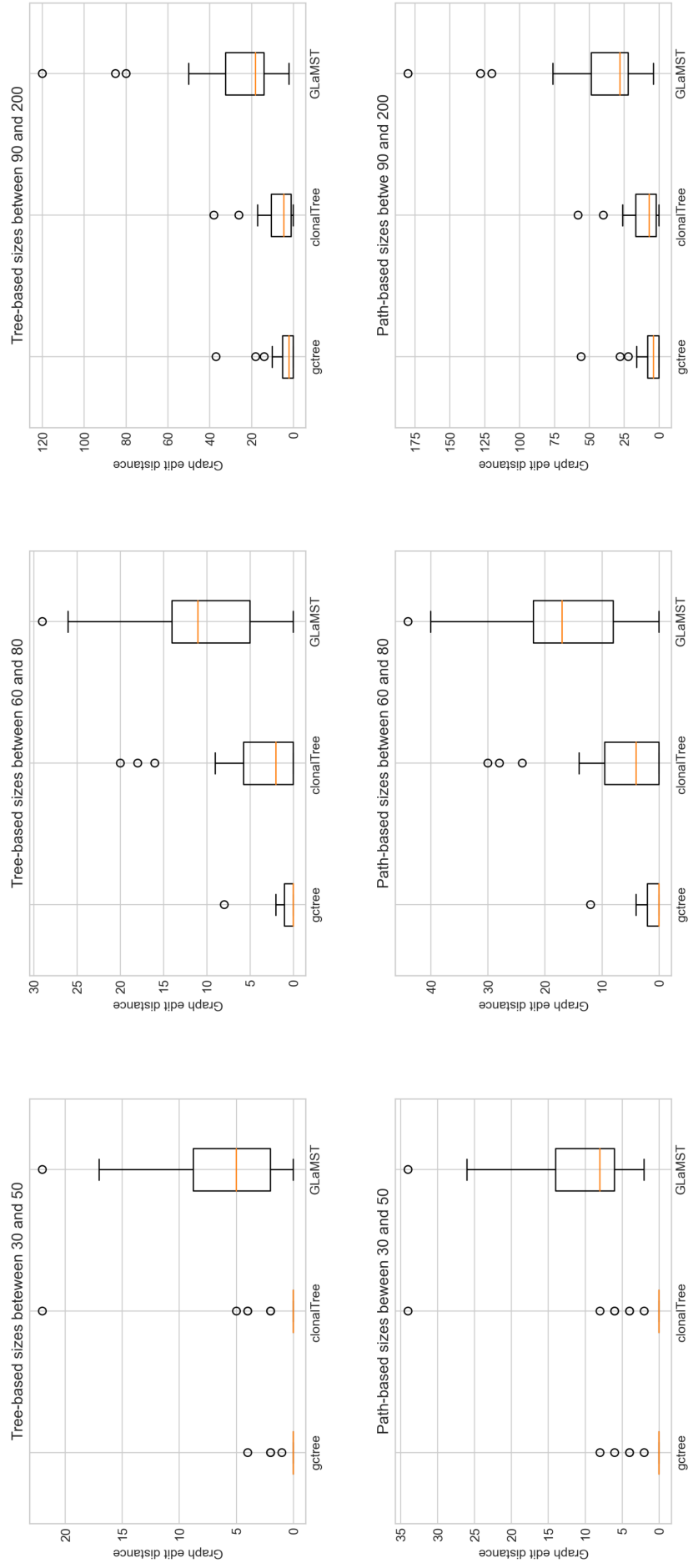


Figure 1: Performance comparison among GCtree, ClonalTree, and GLaMST using GED distances on three categories of trees. The categories are based on the tree sizes.

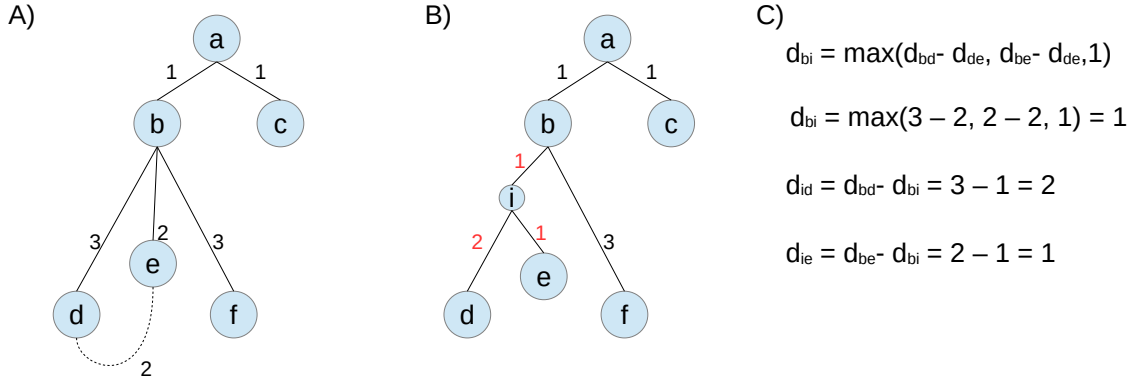


Figure 2: **Editing the reconstructed B cell lineage tree by adding unobserved internal nodes.** Unobserved internal nodes might represent unobserved sequences that were not sampled or disappeared during the affinity maturation. When the distance between two sister nodes is smaller or equal to the distance to their parent (dotted line in A), we add an unobserved internal node as the common ancestor of the two sister nodes (node i in B). C shows how to update edge weights (in red).

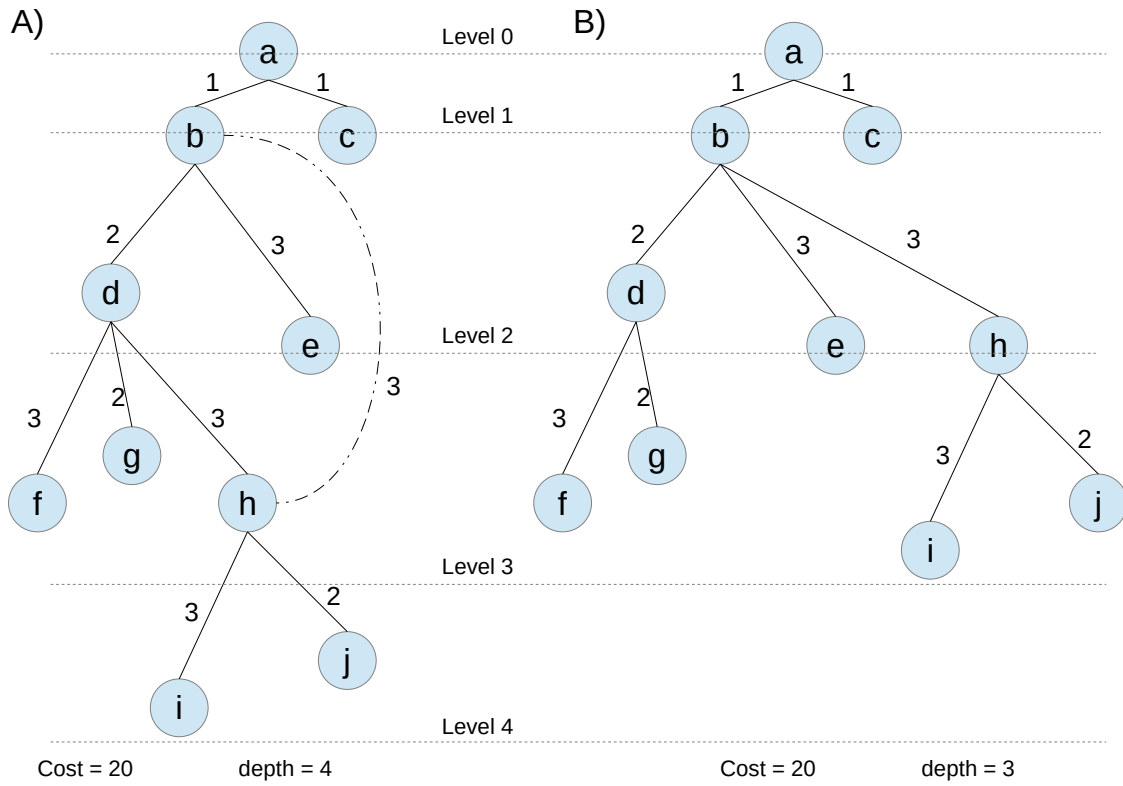
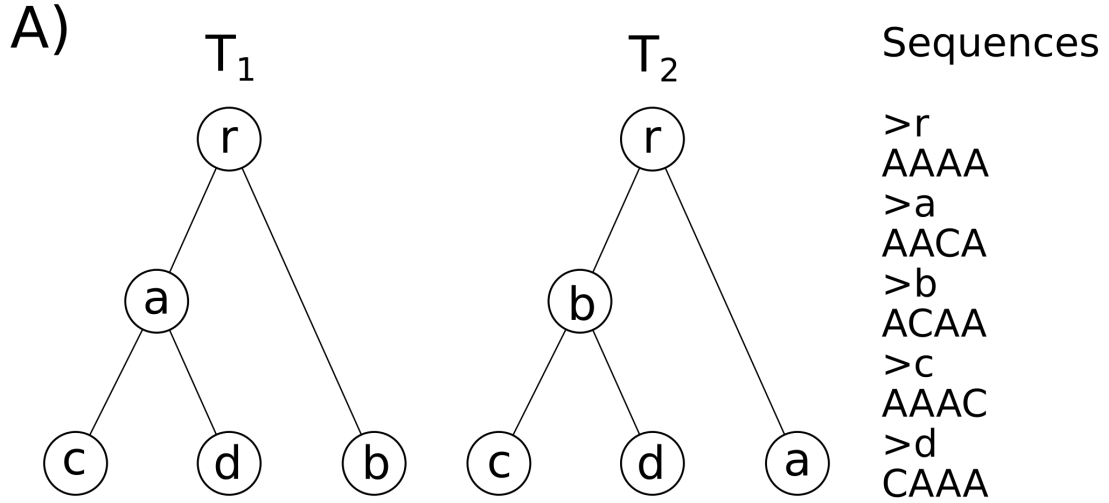


Figure 3: **Editing the reconstructed B cell lineage tree to reduce the depth of the tree while keeping its overall cost.** In this example, the total cost of the tree, or the sum of edge weights, remained the same while the depth of it has been reduced. Edge weights represent the Hamming distance between sequences.



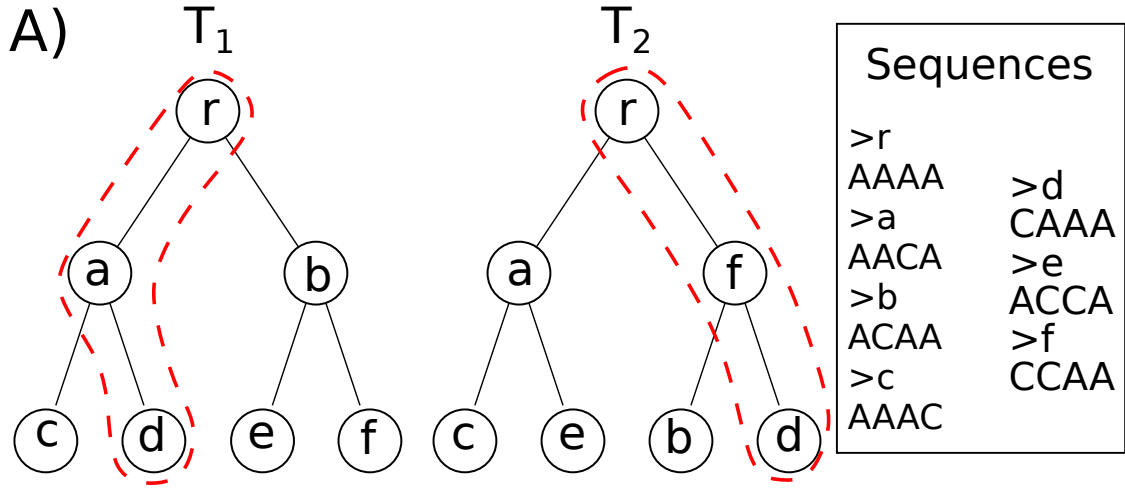
B)

Node pair(i,j)	Ancestral in T_1	Ancestral in T_2	MRCA(i,j)
c, d	a	b	2/4
c, a	a	r	1/4
c, r	r	r	0
c, b	r	b	1/4
d, a	a	r	1/4
d, b	r	b	1/4
d, r	r	r	0
a, r	r	r	0
a, b	r	r	0
b, r	r	r	0
$\sum MRCA(i, j)$			6/4

C)

$$MRCA(T_1, T_2) = \frac{\sum MRCA(i, j)}{C_2^5} = \frac{6/4}{10} = 0.15$$

Figure 4: **An example of MRCA calculation.** A) Let's T_1 and T_2 be two comparable trees, and Sequences a fasta file containing nucleotide sequences associated to each node in both trees. B) For each pair of nodes $(i, j) \in T_1$ and $(i, j) \in T_2$, we find its most recent ancestral in both tree, and computed $MRCA(i, j)$ as the normalized hamming distance between ancestral sequences. C) $MRCA(T_1, T_2)$ is the average of $MRCA(i, j)$ of all pair of nodes.



B)

$p_i = d a r$

$p_j = d f r$

Scoring matrix

		0	1	2
		d	a	r
0	d	0	-2	-1
1	f	-1	-3	-2
2	r	-1	-1	0

C) $C_{i,j} = \max\{(C_{i-1,j} + GT), (C_{i,j-1} + GL), (C_{i-1,j-1} + M_{i-1,j-1})\}$

Dynamic programming matrix

$GT = -inf$ $GL = 0$

	0	d	a	r
0	0	-inf	-inf	-inf
d	1	-inf	0	-inf
f	2	-inf	-inf	-2
r	3	-inf	-inf	-3

$\begin{cases} \uparrow C_{1,2} + GT = -inf + 0 = -inf \\ \leftarrow C_{2,1} + GL = -inf + 0 = -inf \\ \nwarrow C_{1,1} + M_{1,1} = 0 - 3 = -3 \end{cases}$

Figure 5: **An example of COAR calculation for a given leaf (see the algorithm below)** A) Let's T_1 and T_2 be two comparable trees, and Sequences a fasta file containing nucleotide sequences associated to each node in both trees. Let's dotted paths p_i and p_j represent comparable paths in T_1 , and T_2 , respectively. B) Paths to be compared and the scoring matrix containing negative hamming distances for each pair of sequence $S_i \in p_i$ and $S_j \in p_j$. C) We align paths with Needleman-Wunsch alignment algorithm, by using the scoring matrix computed in B, and special gap penalties: $-inf$ for gap top (GT), and 0 for gap left (GL). Each element of the dynamic programming matrix is computed by the formula $C_{i,j}$, see an example for cell $C_{2,2}$. The COAR for the leaf i is $1 = \frac{\min Score_i}{\min(M)} = \frac{C_{3,3}}{-3} = \frac{-3}{-3}$.