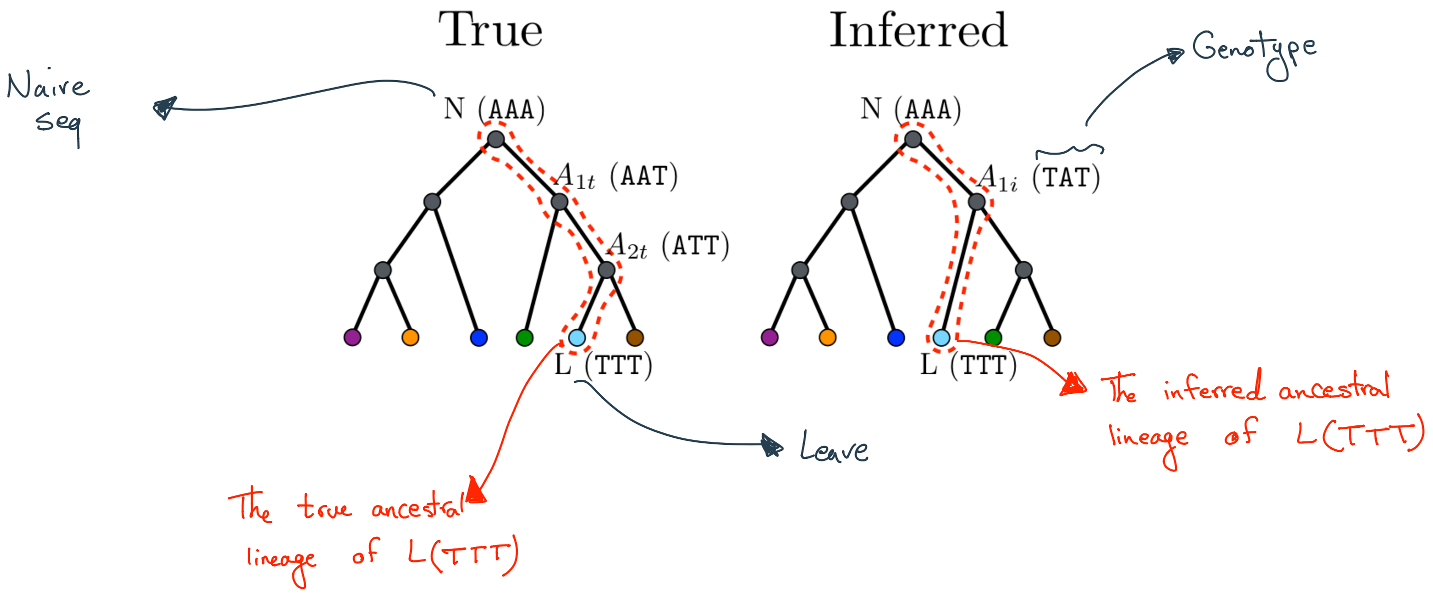
Correctness of ancestral reconstruction

A metric that emphasize the importance of correct ancestral reconstruction and does not penalize the minor topology difference between true and inferred tree when the anstral reconstruction is accurate.



COAR

- compare 2 trees with the same leaves
- For the ancestral lineage i

$$\mathrm{COAR}_i = \frac{\mathrm{alignscore}(\mathrm{leaf}_i)}{\mathrm{alignscore}_{\min}(\mathrm{leaf}_i)}$$

- for the whole tree

$$\mathrm{mean}(\mathrm{COAR}) = \sum_{i=1}^{N_L} \frac{\mathrm{alignscore}(\mathrm{leaf}_i)}{\mathrm{alignscore}_{\min}(\mathrm{leaf}_i)} \Big/ N_L$$

Number of leaves on the tree

The algorithme

1 A lineage = an ordered list

	True	Inferred
Naive (N)	AAA	AAA
A_1	AAT	TAT
A_2	ATT	-
Leaf (L)	TTT	TTT

The list starts with the root and ends with the leaf. (we don't count the start and end elements for the COAR)

2 Creat the score matrix of all pairwise comparison

	N	A_{1t}	A_{2t}	L
N	0	-1	-2	-3
A_{1i}	-2	-1	-2	-1
L	-3	-2	-1	0

→ negative Hamming distance (or other score fonction)

3 Initializing the alignment grid

	-	N	A_{1t}	A_{2t}	L
-	0	-Inf	-Inf	-Inf	-Inf
N	-Inf	→			
A_{1i}	-Inf				
L	-Inf				

. We use -Inf to force the alignment to start at the root.

We do not allow gap in the longest list (for unidentical true vs inferred list)

4 Fill the matrix

$$C_{i,j} = \max\left\{(C_{i-1,j} + gp_{\text{down}}); (C_{i,j-1} + gp_{\text{right}}); (C_{i-1,j-1} + \overbrace{S_{i-1,j-1})}^{\text{from the score matrix}}\right\}$$

We use this formula

In this example the true ancestral list is the longer one, so we don't allow the gap in it.

So we use : $gp_{\text{down}} = -\text{Inf}$ and $gp_{\text{right}} = 0$.

	-	N	A_{1t}	A_{2t}	L
-	0	-Inf	-Inf	-Inf	-Inf
N	-Inf	0	0	0	0
A_{1i}	-Inf	-Inf	-1	-1	-1
L	-Inf	-Inf	-Inf	-2	-1

5 The traceback

$$\text{move}_{i,j} = \text{which}\,\{C_{i,j} = [(C_{i-1,j} + gp_{\text{down}}), (C_{i,j-1} + gp_{\text{right}}), (C_{i-1,j-1} + S_{i-1,j-1})]\}$$

6 Finding the alignment, calculat the penality, normalize it by max penality and calculate COAR

True	N	A_{1t}	A_{2t}	L
Inferred	N	A_{1i}	-	L
Penalty	0	-1	0	0
Max penalty	0	-3	0	0
COAR	-1/-3=0.333			

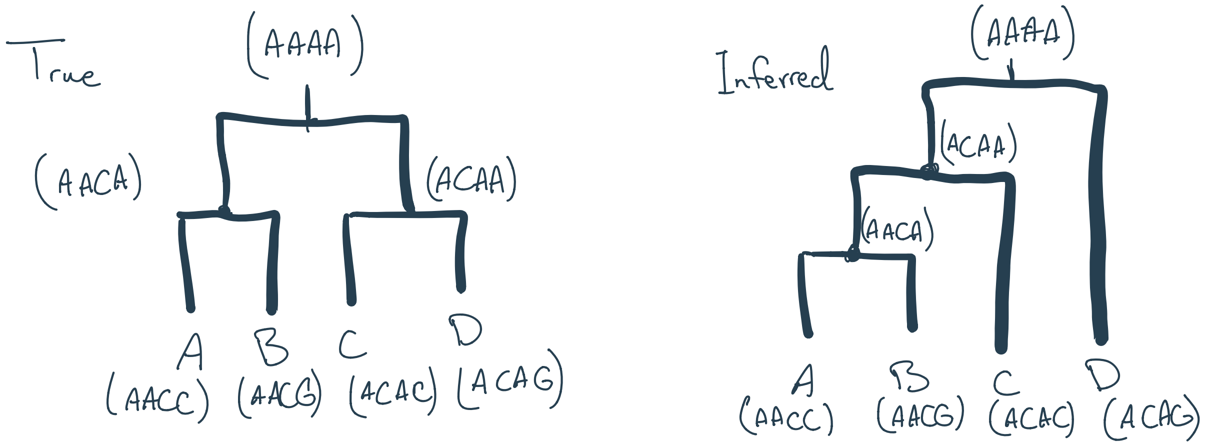
No similarity between 2 lists

bad ancestral reconstruction

$0 < \text{COAR} < 1$

perfect ancestral reconstruction

Metric using most recent common ancester (MRCA)



for a given pair of leaves in the tree the MRCA is :
The average Hamming distance between the true and the inferred ancestral seq

for comparing 2 trees
- we have to iterat over all combination of leaves pair

$$\sum_{i=1}^{N}\sum_{j=i+1}^{N} d_H(\underbrace{T_{i,j}}_{\text{MRCA of i and j on the true tree}}, \overbrace{I_{i,j}}^{\text{MRCA of i and j on the inferred tree}}) \Big/ (\underbrace{N}_{\text{number of seq}}(N-1)/2)\overset{\text{seq length}}{L}.$$

	AB	AC	AD	BC	BD	CD
T	AACA	AAAA	AAAA	AAAA	AAAA	ACAA
I	AACA	ACAA	AAAA	ACAA	AAAA	AAAA
dH	0	$\frac{1}{4}$	0	$\frac{1}{4}$	0	$\frac{1}{4}$

$$\Rightarrow \left(\frac{1}{4}\right) \times \cancelto{1}{3} \times \frac{1}{\cancelto{2}{6}}$$

$$= \frac{1}{}$$