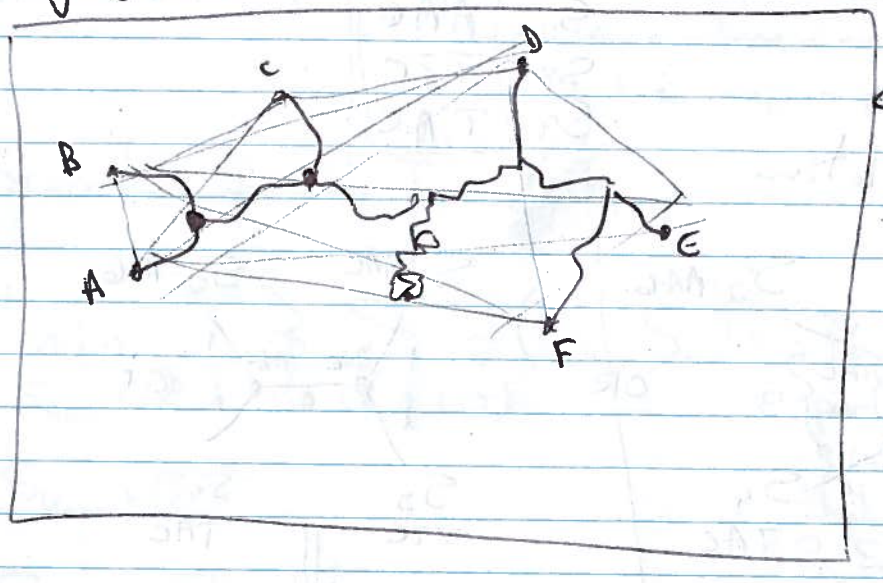
Phylogenetic Inference - Maximum Parsimony



Large Parsimony Problem:

Given: $S_1 \ldots S_n$ of length L, in multiple sequence aln, with columns containing gaps removed

Find:
- Tree T with leaves labeled with $S_1 \ldots S_n$
- Ancestral sequence S_u for each internal node of T

such that $\sum_{(u,v) \in E(T)} d(S_u, S_v)$ is minimized

$d(S_u, S_v)$ = # of substitutions btw S_u, S_v

Example: Input:

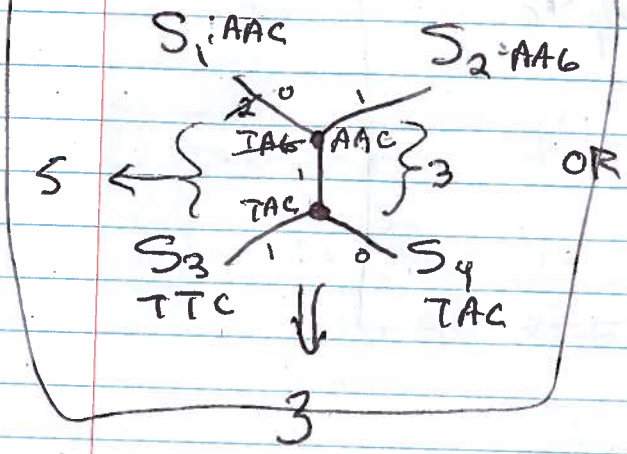
MSA

S_1: AAC
S_2: AA6
S_3: TTC
S_4: TAC

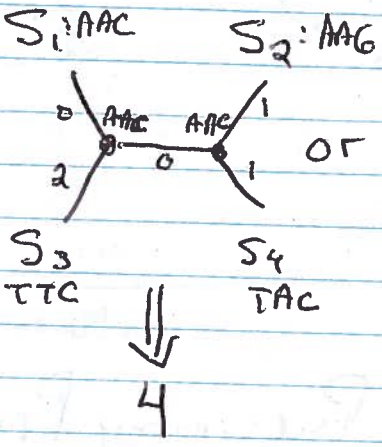
Unrooted tree

Rooted tree

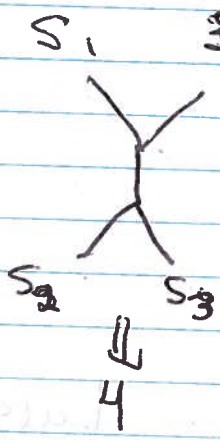
Opt-Solution



3

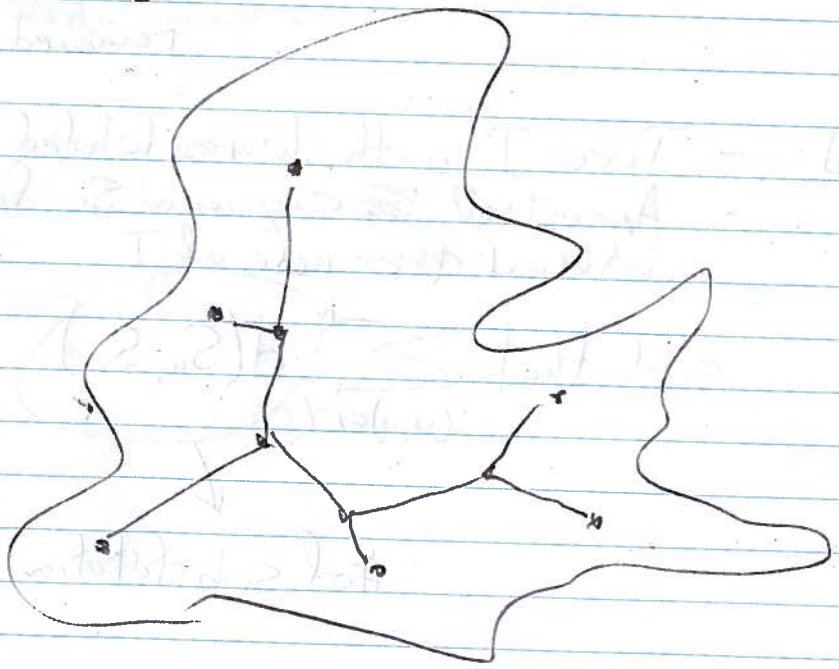


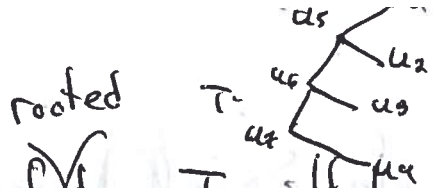
4



4

Steiner tree





Define Parsimony Score of tree T with leaves $u_1, u_2 \dots u_n$ and internal nodes $u_{n+1}, u_{n+2} \dots, u_{2n-1}$

$$\text{ParsScore}(S_1, S_2 \dots S_n, T) = \min_{\substack{S_{u_{n+1}} \in (\Sigma)^n \\ S_{u_{n+2}} \in (\Sigma)^n \\ \vdots \\ S_{u_{2n-1}} \in \Sigma^n}} \left\{ \sum_{(u,v) \in E(T)} d(S_u, S_v) \right\}$$

$\equiv$ The minimum # of substitutions performed along branches of T that can explain $S_1, S_2 \dots S_n$

$\left(4^L\right)^{n-1}$ cases $= 4^{L(n-1)}$

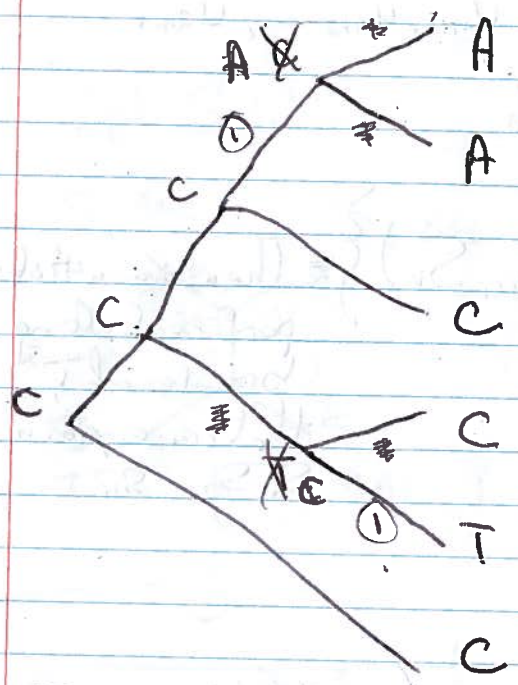
Observation: $\text{ParsScore}(S_1 \dots S_n, T) = \sum_{i=1}^{L} \text{ParsScore}(S_1[i], \dots, S_n[i], T)$

→ Small Parsimony Problem

Given: - Aligned $S_1, \dots, S_n$ of length L
- Phylogenetic tree T

Goal: Calculate $\text{ParsScore}(S_1 \dots S_n, T)$

Example: Small Parsimony Problem

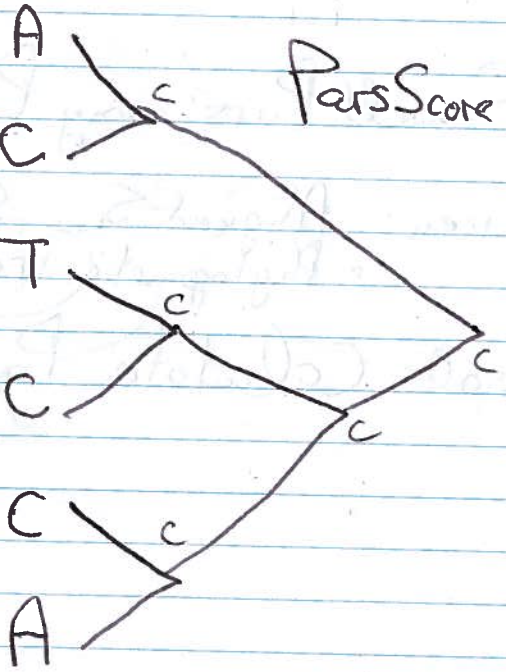


Score = 4

Score = 2 Optimal sol.

$\Rightarrow$ ParsScore(A, A, C, C, C, T, C, Tree) = 2

How to calculate $ParsScore[S_1[i], S_2[i], \ldots S_n[i], T]$?



ParsScore = 3