Distance based phylogeny reconstruction

Usman Roshan

Phylogenetics

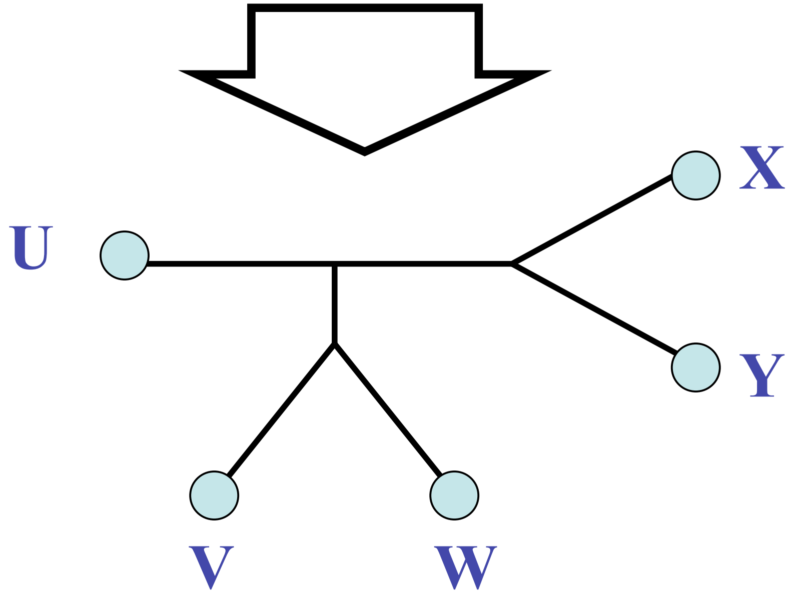
- Study of how species relate to each other
- “Nothing in biology makes sense, except in the light of evolution”, Theodosius Dobzhansky, *Am. Biol. Teacher* (1973)
- Rich in computational problems
- Fundamental tool in comparative bioinformatics

Why phylogenetics?

- Study of evolution
  - Origin and migration of humans
  - Origin and spead of disease
- Many applications in comparative bioinformatics
  - Sequence alignment
  - Motif detection (phylogenetic motifs, evolutionary trace, phylogenetic footprinting)
  - Correlated mutation (useful for structural contact prediction)
  - Protein interaction
  - Gene networks
  - Vaccine devlopment
  - And many more…

Phylogeny Problem

U	V	W	X	Y
AGGGCAT	TAGCCCA	TAGACTT	TGCACAA	TGCGCTT



Phylogeny Problem

- Two main methodologies:
  - Alignment first and phylogeny second
    - Construct alignment using one of the MANY alignment programs in the literature
    - Do manual (eye) adjustments if necessary
    - Apply a phylogeny reconstruction method
    - Fast but biologically not realistic
    - Phylogeny is highly dependent on accuracy of alignment (but so is the alignment on the phylogeny!)
  - Simultaneously alignment and phylogeny reconstruction
    - Output both an alignment and phylogeny
    - Computationally much harder
    - Biologically more realistic as insertions, deletions, and mutations occur during the evolutionary process

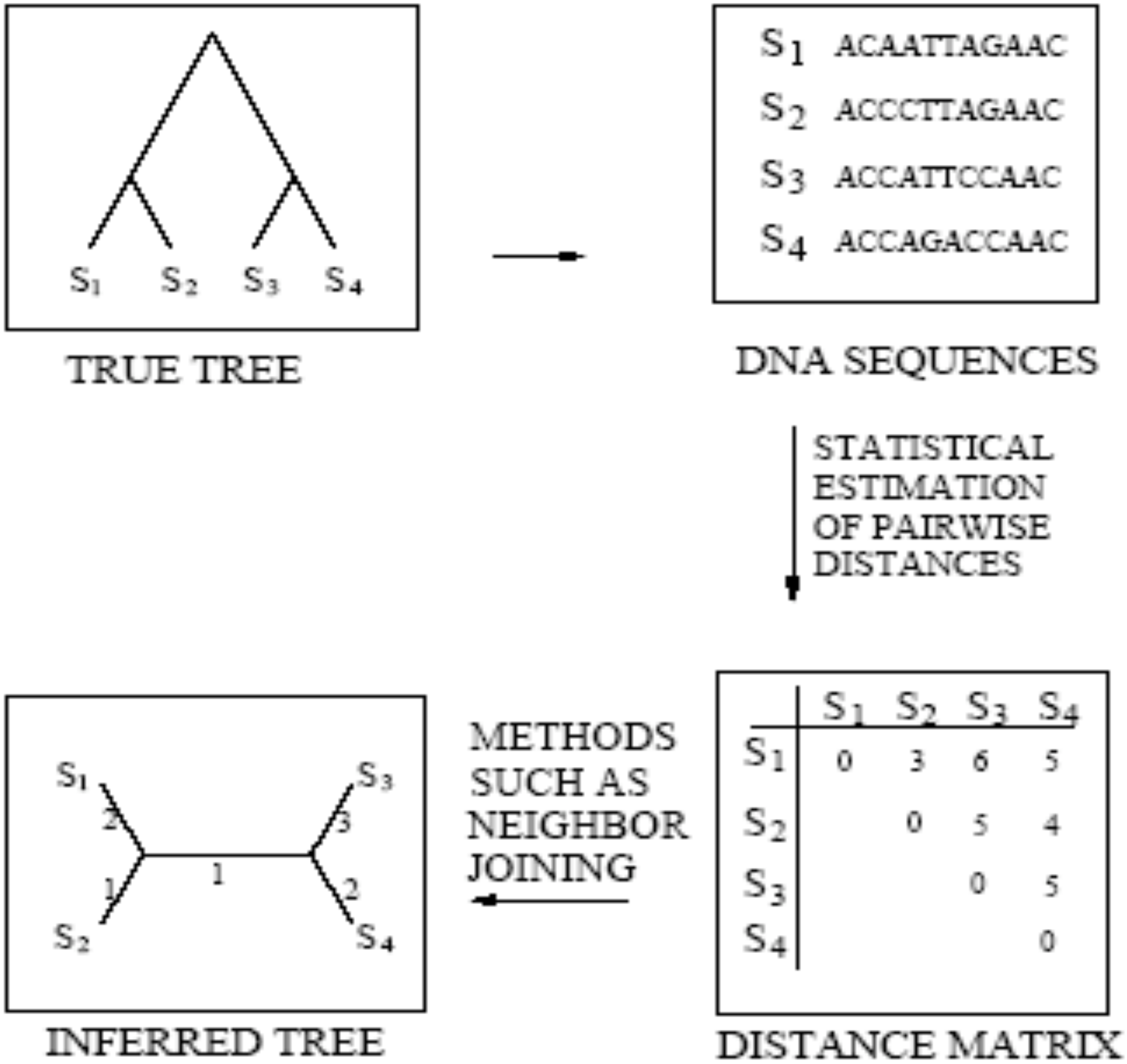
First methodology

- Compute alignment (for now we assume we are given an alignment)
- Construct a phylogeny (two approaches)
- Distance-based methods
  - Input: Distance matrix containing pairwise statistical estimation of aligned sequences
  - Output: Phylogenetic tree
  - Fast but less accurate
- Character-based methods
  - Input: Sequence alignment
  - Output: Phylogenetic tree
  - Accurate but computationally very hard

Definitions

- Tree:
  - Set of nodes and edges
  - Undirected graph
  - No cycles
  - Connected
- Examples
- Degree of node = number of edges connected to the node
- Binary tree: every node has at most two children
- Phylogeny: unrooted binary tree

Distance-based methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

DNA SEQUENCES

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

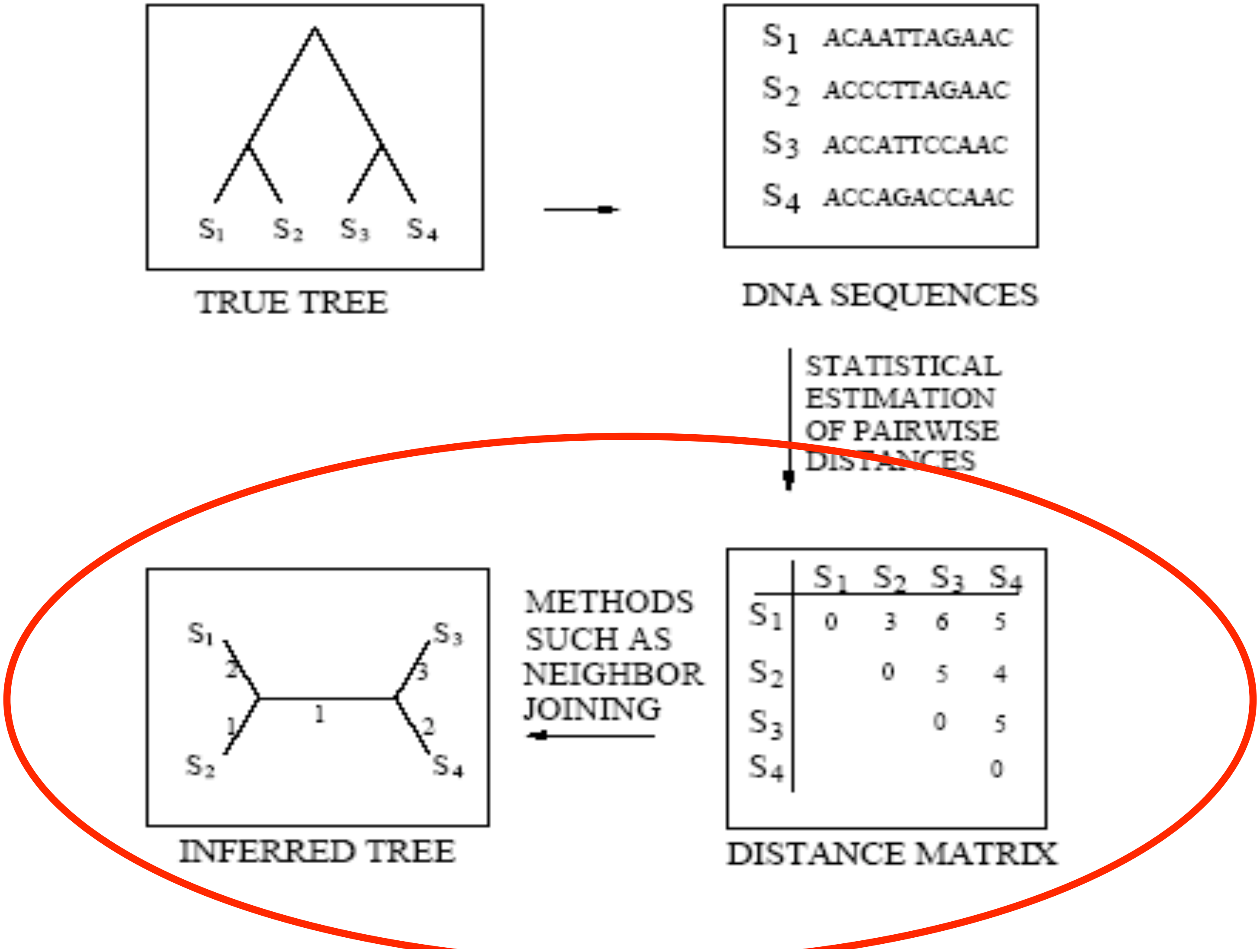
	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING

INFERRED TREE

Distance methods



S_1 ACAATTAGAAC
S_2 ACCCTTAGAAC
S_3 ACCATTCCAAC
S_4 ACCAGACCAAC

DNA SEQUENCES

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING

Distance methods

- UPGMA: similar to hierarchical clustering but not additive
- Neighbor-joining: more sophisticated and additive
- What is additivity?

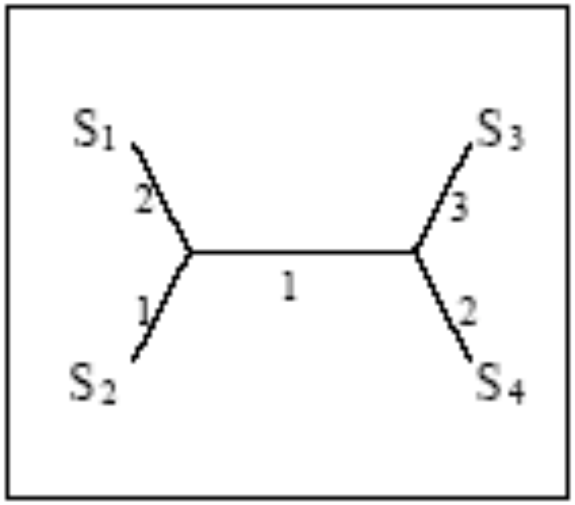
Additivity

A distance matrix D is additive if there exists a tree, $T = (V, E)$, and $w : E \rightarrow \mathcal{R}^+$ such that $D_{ij} = \sum_{e \in P_{ij}} w(e)$.

Waterman *et al*, 1977, showed that:

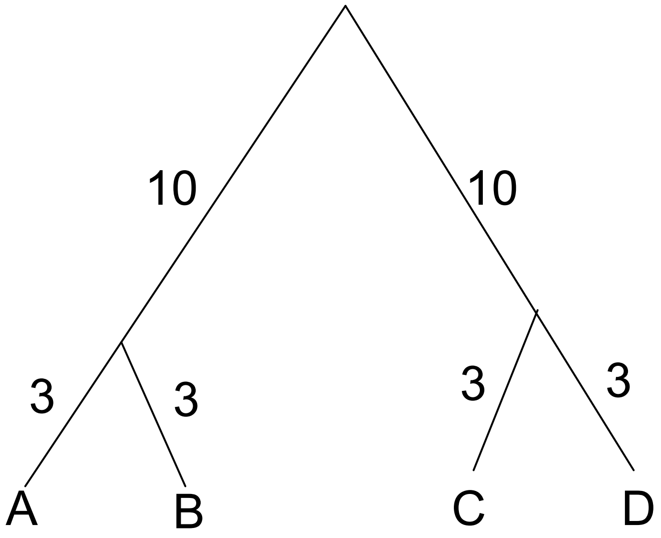
	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

POLYTIME
INVERTIBLE
⟷



UPGMA

UPGMA is not additive but works for ultrametric trees. Takes *O(n^3)* time



	A	B	C	D
A		6	26	26
B			26	26
C				6
D				

UPGMA

Input: distance matrix *D*, **Output**: Phylogeny *T*

1. Set *d=D*
2. Initialize *n* clusters where each cluster *i* contains the sequence *I*
3. Find closest pair of clusters *i, j,* using distances in matrix *d*
4. Make them neighbors in the tree by adding new node *(ij),* and set distance from *(ij)* to *i* and *j* as $d_{ij}/2$
5. Update distance matrix *d* with weighted average. For all clusters *k* do the following (*ni* and *nj* are size of clusters *i* and *j* respectively)

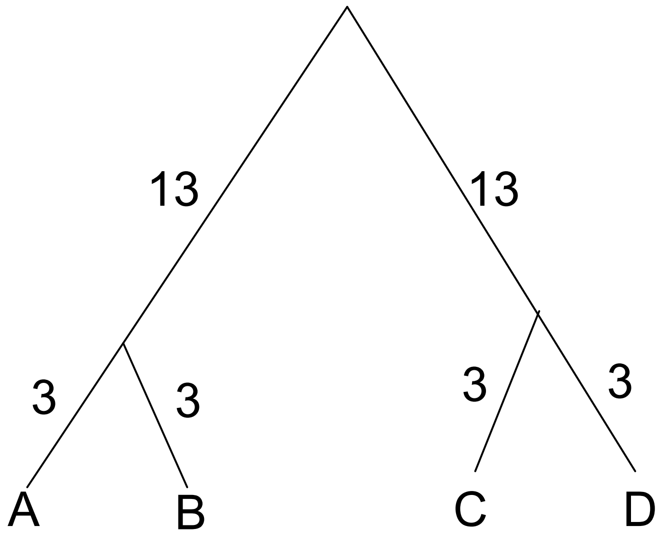
$$d(i,j) = \frac{1}{n_i n_j} \sum_{i' \in L(i)} \sum_{j' \in L(j)} D(i',j')$$

UPGMA

6. Delete columns and rows for *i* and *j* in *d* and add new ones corresponding to cluster *(ij)* with distances as computed above
7. Goto step 2 until only one cluster is left

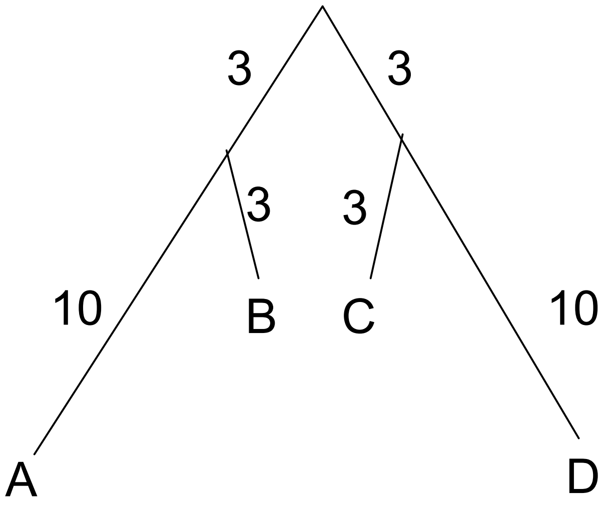
UPGMA

	A	B	C	D
A		6	32	32
B			32	32
C				6
D				



UPGMA

Doesn’t work (in general) for non ultrametric trees

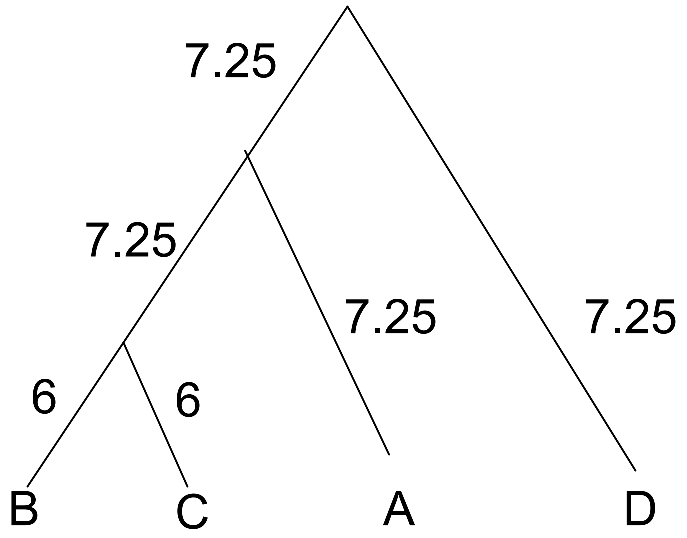


	A	B	C	D
A		13	19	26
B			12	19
C				13
D				

UPGMA

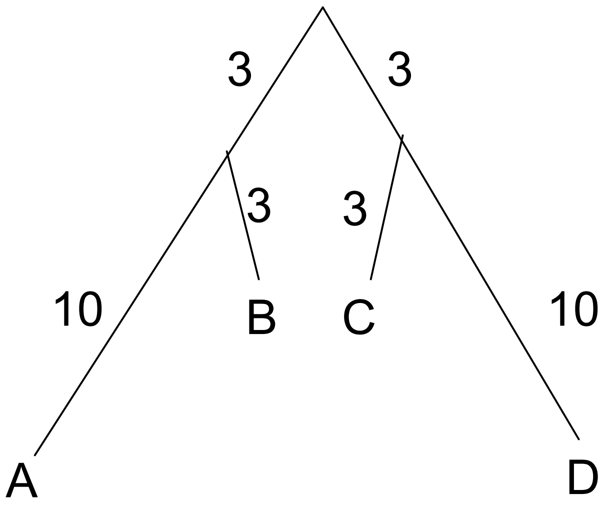
UPGMA constructs incorrect tree here

	A	B	C	D
A		13	19	26
B			12	19
C				13
D				

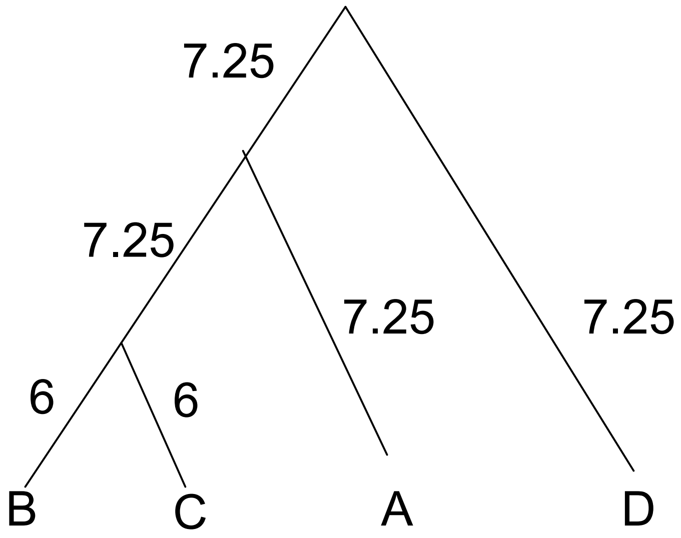


UPGMA

Bipartition (*BC,AD)* is not in true tree



True tree



UPGMA tree