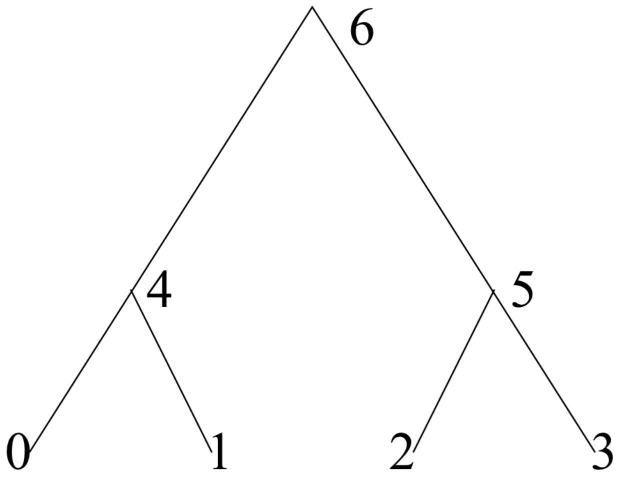
UPGMA implementation

Usman Roshan

Representing phylogenies

- We use three lists to represent rooted binary trees: parent, left child, and right child.
- The node id is the index in the array.
- For leaves the left and right child is set to -1.
- If a node has no parent then its value in the list is set to -1.

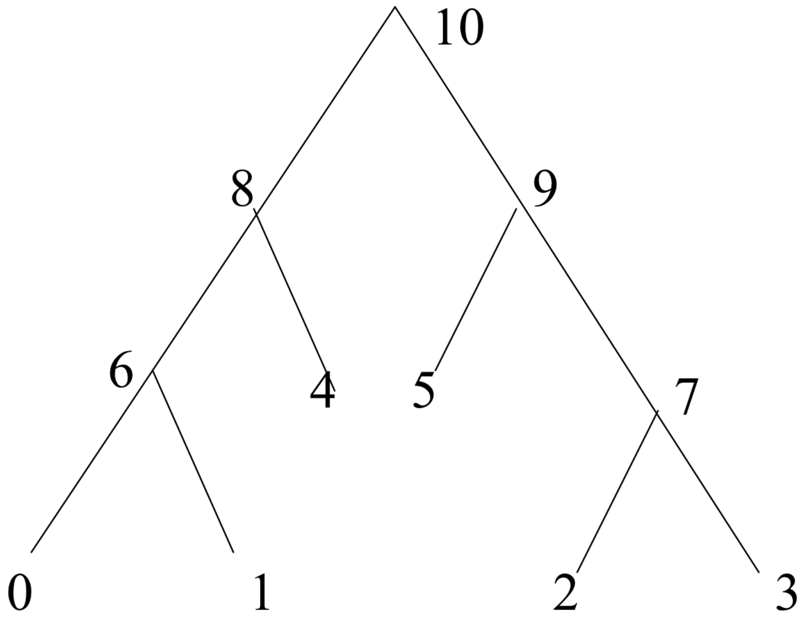
Representing phylogenies



	0	1	2	3	4	5	6
P:	4	4	5	5	6	6	-1
L:	-1	-1	-1	-1	0	2	4
R:	-1	-1	-1	-1	1	3	5

Representing phylogenies

- Write the P, L, and R lists for the phylogeny below.



UPGMA pseudocode

- Read distance matrix D from file and make copy of it called d.
- Initialize P, L, and R lists
- While (nodes_to_do > 1):
  - Find closest pair in distance matrix d
  - Add new node in tree
  - Update distance matrix
  - nodes_to_do = nodes_to_do - 1

UPGMA pseudocode

- Initialization:
  - D has dimensions n x n where n is the number of sequences (leaves)
  - The final tree will have 2n + 1 nodes
  - We set d to be of dimension 2n + 1
- Find closest pair in d:
  - Returns indices i and j that are closest
  - Ignore entries that are non-positive
- Add new node in tree
  - Parameters are P, L, R, new_node_id and indices i and j from above function

UPGMA pseudocode

- Update distance matrix d
  - Calculate distances from new node to nodes that have parent set to -1
  - Set row and column of nodes i and j (that were returned by find_closest_pair) to -1