BIOINFORMATICS AND STATISTICAL GENETICS

GABRIEL VALIENTE

ALGORITHMS, BIOINFORMATICS, COMPLEXITY AND FORMAL METHODS RESEARCH GROUP,
TECHNICAL UNIVERSITY OF CATALONIA

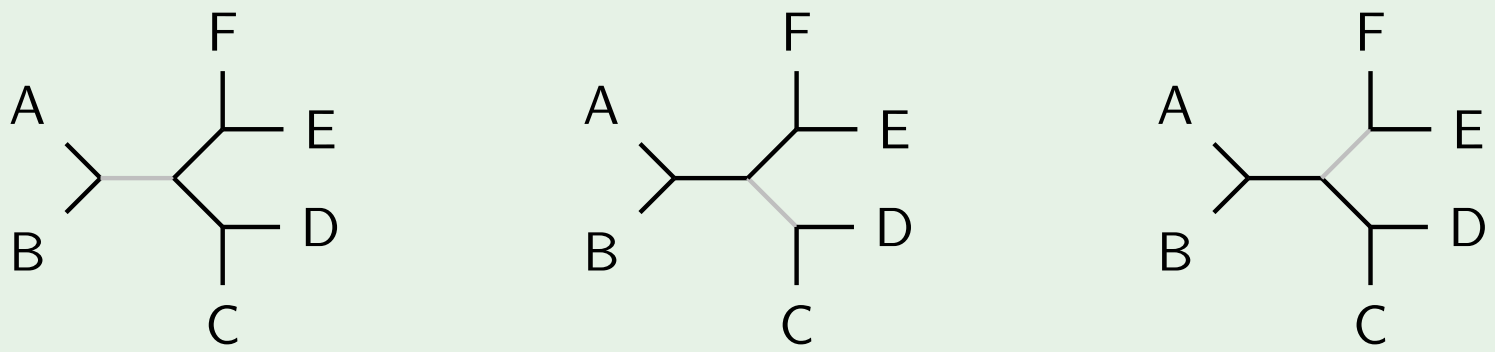
2023–2024

Agreement of phylogenetic trees

- Partition distance
- Nodal distance
- Triplets distance
- Transposition distance
- Edit distance
- Alignment of phylogenetic trees

- The partition distance is based on the partition of the taxa induced by each internal branch in the two phylogenetic trees under comparison.
- While cutting a tree along each of the the external branches partitions the set of taxa in a trivial way (with each partition consisting of a single taxon on one side and the remaining taxa on the other side), cutting along each of the internal branches reveals similarities and differences between the two trees.

Example



- For rooted phylogenetic trees, the partition of the taxa induced by an internal branch (v, w) consists of the labels of all terminal nodes which are descendants of node w and the labels of all other terminal nodes.

- The partition distance between two rooted phylogenetic trees can be computed by first obtaining, for each of the two phylogenetic trees, the partition of the taxa induced by each of their internal branches and then counting the number of partitions of the taxa in each of the trees that do not belong to the partitions of the taxa in the other tree.

function partition(T)
  $P \leftarrow \emptyset$
  for each internal node v of T **do**
    $A \leftarrow$ descendants of v in T
    $B \leftarrow$ all other leaves of T
    $P \leftarrow P \cup \{(A, B)\}$
  return P

function partition distance(T_1, T_2)
  $P_1 \leftarrow partition(T_1)$
  $P_2 \leftarrow partition(T_2)$
  $d \leftarrow 0$
  for $(A, B) \in P_1$ **do**
    if $(A, B) \notin P_2$ **then**
      $d \leftarrow d + 1$
  for $(A, B) \in P_2$ **do**
    if $(A, B) \notin P_1$ **then**
      $d \leftarrow d + 1$
  return d

- Let $C(T)$ be the set of clusters (set of sets of descendent node labels) in T
- RF distance

$$RF(T_1, T_2) = \frac{|C(T_1) \oplus C(T_2)|}{|C(T_1) \cup C(T_2)|}$$

- The RF distance is a metric for phylogenetic trees
- The RF distance can be computed in $O(n)$ time

- D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1–2):131–147, 1981
- W. H. E. Day. Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification*, 2(1):7–28, 1985

- Let $C(T)$ be the set of clusters (set of sets of descendent node labels) in T
- GRF distance

$$GRF(T_1, T_2) = \frac{\sum_{x \in C(T_1)} \sum_{y \in C(T_2) \setminus C(T_1)} |x \oplus y|}{|C(T_1) \cup C(T_2)| \cdot |C(T_1)|} + \frac{\sum_{x \in C(T_1) \setminus C(T_2)} \sum_{y \in C(T_2)} |x \oplus y|}{|C(T_1) \cup C(T_2)| \cdot |C(T_2)|}$$

- The GRF distance is a metric for phylogenetic trees
- The GRF distance can be computed in $O(n)$ time

- M. Llabrés, F. Rosselló, and G. Valiente. A generalized Robinson-Foulds distance for clonal trees, mutation trees, and phylogenetic trees and networks. In *Proc. 11th ACM Int. Conf. Bioinformatics, Computational Biology and Health Informatics*, pages 13:1–13:10, New York, NY, 2020. ACM Press
- M. Llabrés, F. Rosselló, and G. Valiente. The generalized Robinson-Foulds distance for phylogenetic trees. *Journal of Computational Biology*, 28(12):1–15, 2021

- The nodal distance, also called path difference metric, is based on the distances between each two terminal nodes in the two trees under comparison.
- Let $D(T)$ be the length $n(n-1)/2$ vector of nodal distances between each pair of terminal nodes of an unrooted phylogenetic tree T, that is,

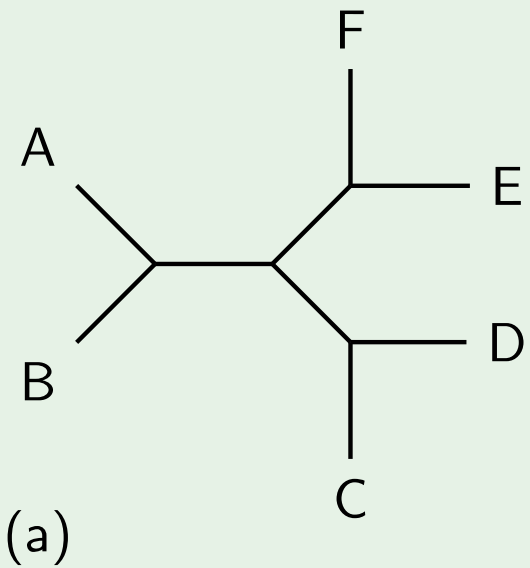
$$D(T) = (d_T(1,2), d_T(1,3), \ldots, d_T(1,n), d_T(2,3), \ldots, d_T(n-1,n)),$$

  where the n terminal nodes of T are numbered $1, \ldots, n$.
- The nodal distance $d_N(T_1, T_2)$ between two unrooted phylogenetic trees T_1 and T_2 is the sum of the absolute differences between their vectors of nodal distances, that is,

$$d_N(T_1, T_2) = \sum_{\substack{1 \leqslant i < n \\ i < j \leqslant n}} |d_{T_1}(i,j) - d_{T_2}(i,j)|$$

- J. Bluis and D. G. Shin. Nodal distance algorithm: Calculating a phylogenetic tree comparison metric. In *Proc. 3rd IEEE Symp. BioInformatics and BioEngineering*, pages 87–94. IEEE Computer Society, 2003

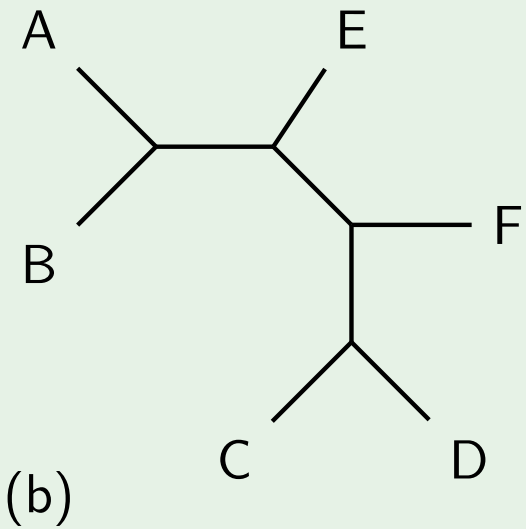
Example



(a)

	A	B	C	D	E	F
A	0	2	4	4	4	4
B	2	0	4	4	4	4
C	4	4	0	2	4	4
D	4	4	2	0	4	4
E	4	4	4	4	0	2
F	4	4	4	4	2	0

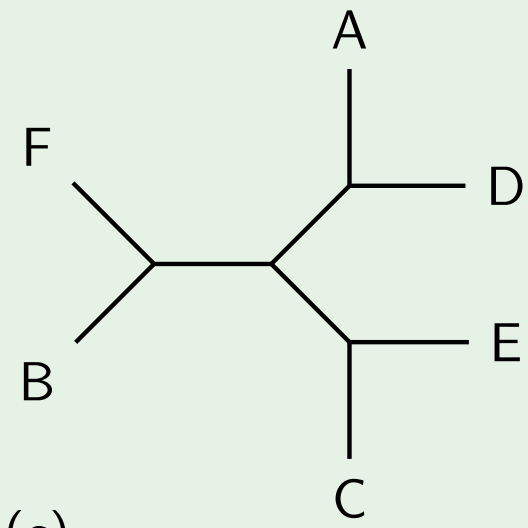
Example



(b)

	A	B	C	D	E	F
A	0	2	5	5	3	4
B	2	0	5	5	3	4
C	5	5	0	2	4	3
D	5	5	2	0	4	3
E	3	3	4	4	0	3
F	4	4	3	3	3	0

Example



(c)

	A	B	C	D	E	F
A	0	4	4	2	4	4
B	4	0	4	4	4	2
C	4	4	0	4	2	4
D	2	4	4	0	4	4
E	4	4	2	4	0	4
F	4	2	4	4	4	0

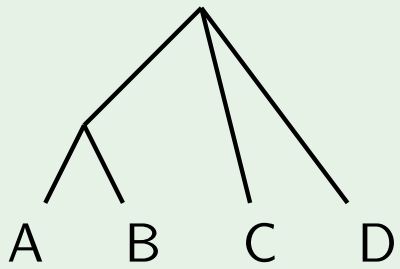
Example

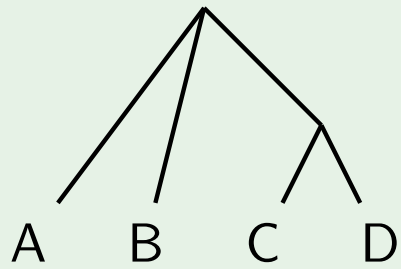
	(a)	(b)	(c)	$\lvert(a)-(b)\rvert$	$\lvert(a)-(c)\rvert$	$\lvert(b)-(c)\rvert$
AB	2	2	4	0	2	2
AC	4	5	4	1	0	1
AD	4	5	2	1	2	3
AE	4	3	4	1	0	1
AF	4	4	4	0	0	0
BC	4	5	4	1	0	1
BD	4	5	4	1	0	1
BE	4	3	4	1	0	1
BF	4	4	2	0	2	2
CD	2	2	4	0	2	2
CE	4	4	2	0	2	2
CF	4	3	4	1	0	1
DE	4	4	4	0	0	0
DF	4	3	4	1	0	1
EF	2	3	4	1	2	1
				9	12	19

- When the phylogenetic trees are rooted and not binary, the nodal distance fails to be a metric on the space of rooted phylogenetic trees.
- For instance, the following two rooted phylogenetic trees have nodal distance zero, but they are non-isomorphic.

Example

The rooted phylogenetic trees with Newick string `((A,B),C,D);` (left) and `(A,B,(C,D));` (right) have the same vectors of nodal distances and, thus, their nodal distance is 0.





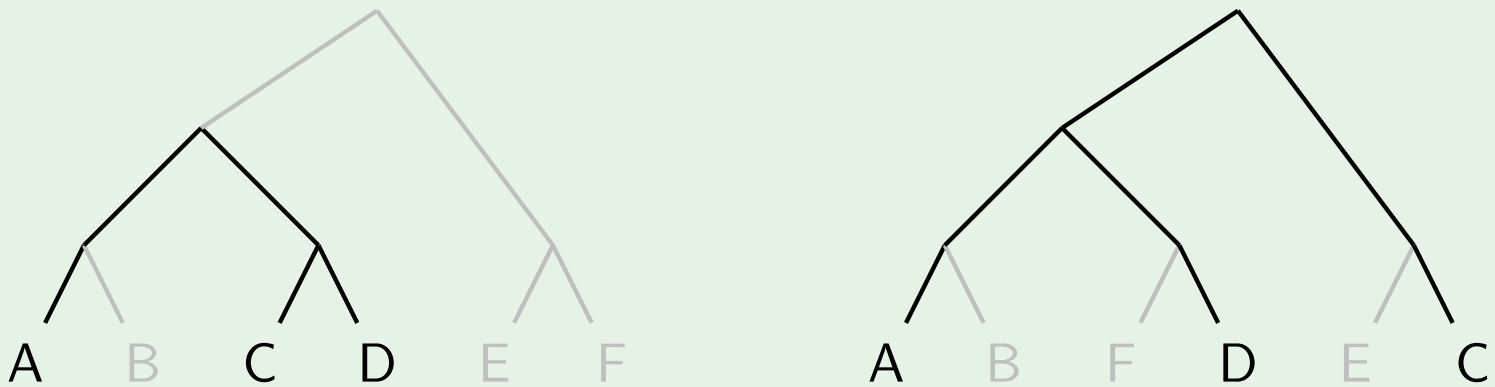
- The nodal distance is a metric for unrooted phylogenetic trees
- The nodal distance is a metric for rooted binary phylogenetic trees
- The nodal distance can be computed in $O(n^2)$ time

- The splitted nodal distance is a metric for phylogenetic trees
- The splitted nodal distance can be computed in $O(n^2)$ time

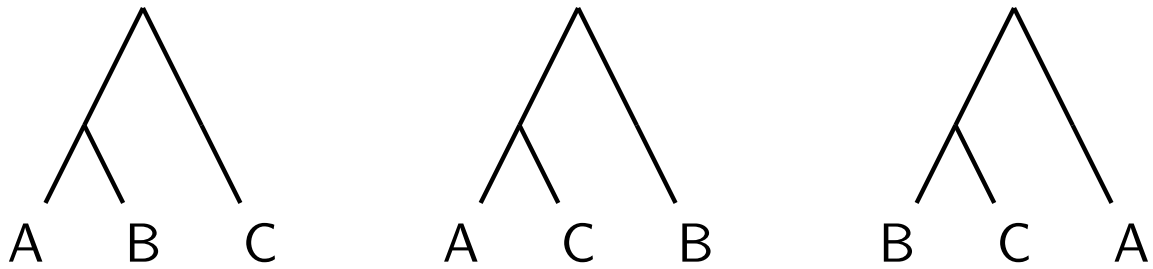
- G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente. Nodal distances for rooted phylogenetic trees. *Journal of Mathematical Biology*, 61(2):253–276, 2010

- The triplets distance is based on the subtrees induced by triplets of terminal nodes in the two phylogenetic trees under comparison.
- The sets of subtrees induced by triplets of terminal nodes reveal similarities and differences between two rooted binary phylogenetic trees.
- The triplets distance between two binary phylogenetic trees labeled over the same taxa is defined as the size of the symmetric difference of their sets of triplets, that is, the number of triplets in which the two phylogenetic trees differ.

Example



- A rooted phylogenetic tree can be reconstructed in a unique way from the set of all its triplet topologies.
- Given a triplet of terminal nodes, there are only three possible induced subtrees if the phylogenetic tree is binary.



function triplet(T, i, j, k)
    $ij \leftarrow LCA(T, i, j)$
    $ik \leftarrow LCA(T, i, k)$
    $jk \leftarrow LCA(T, j, k)$
    if $ik = jk$ **then**
        return $((i, j), k)$;
    else
        if $ij = jk$ **then**
            return $((i, k), j)$;
        else
            return $((j, k), i)$;

- The triplets distance between two phylogenetic trees can be computed by first obtaining the triplet induced by each set of three terminal node labels in each of the trees and then counting the number of triplets in which the two trees differ.

function triplets_distance(T_1, T_2)
    $L \leftarrow$ terminal node labels in T_1 and T_2
    $n \leftarrow length(L)$
    $d \leftarrow 0$
    for $i \leftarrow 1, \ldots, n$ **do**
        for $j \leftarrow i+1, \ldots, n$ **do**
            for $k \leftarrow j+1, \ldots, n$ **do**
                $t_1 \leftarrow triplet(T_1, L[i], L[j], L[k])$
                $t_2 \leftarrow triplet(T_2, L[i], L[j], L[k])$
                if $t_1 \neq t_2$ **then**
                    $d \leftarrow d+2$
    return d

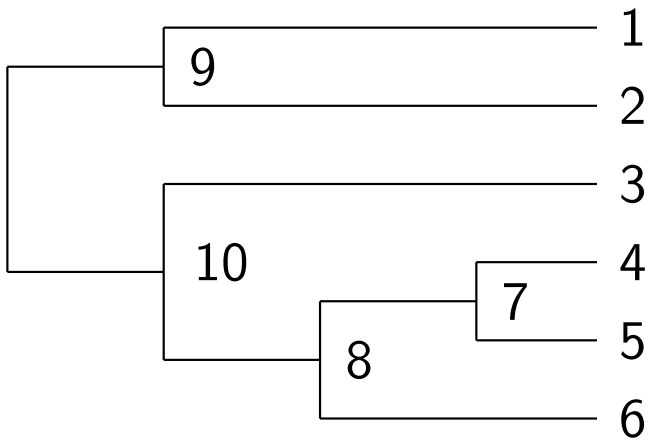
- The triplets distance is a metric for phylogenetic trees
- The triplets distance can be computed in $O(n^3)$ time

- D. E. Critchlow, D. K. Pearl, and C. Qian. The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology*, 45(3):323–334, 1996

- Let $i \prec j$ denote that i is a predecessor of j in sorted bottom-up order
- Let the internal nodes of T be labeled as $\ell(j) = \max\{\ell(i) \mid i \prec j\} + 1$
- The matching representation of a binary phylogenetic tree $T = (V, E)$ is the partition of $\{1, \ldots, 2n\}$ into 2-subsets

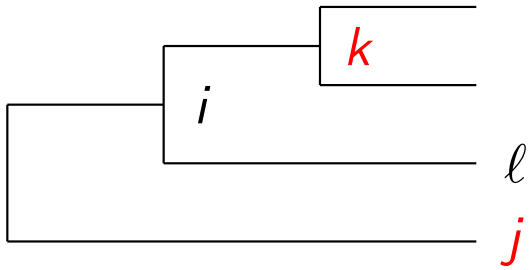
$$M(T) = \{\{\ell(j), \ell(k)\} \mid \exists i : (i, j), (i, k) \in E\}$$

- Example

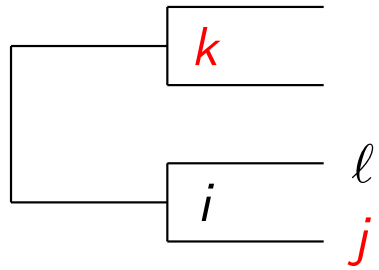


- $M(T) = \{\{1, 2\}, \{3, 8\}, \{4, 5\}, \{6, 7\}, \{9, 10\}\}$

- Let M be a partition of $\{1, \ldots, 2n\}$ into 2-subsets, and let $\{i, j\}, \{k, \ell\} \in M$
- The transposition of M at j and k is the replacement of $\{i, j\}$ by $\{i, k\}$ and $\{k, \ell\}$ by $\{j, \ell\}$ in M



$\{i, j\}, \{k, \ell\}$



$\{i, k\}, \{j, \ell\}$

- The matching distance $MD(T_1, T_2)$ is the minimum number of transpositions needed to transform $M(T_1)$ into $M(T_2)$
- The transposition distance $TD(T_1, T_2)$ is the matching distance $MD(T_1|L, T_2|L)$ between the topological restriction of T_1 and T_2 to their common taxa L
- The transposition distance is a metric for binary phylogenetic trees
- The transposition distance is a metric for phylogenetic trees
- The transposition distance can be computed in $O(n)$ time

- G. Valiente. A fast algorithmic technique for comparing large phylogenetic trees. In *Proc. 12th Int. Symp. String Processing and Information Retrieval*, volume 3772 of *Lecture Notes in Computer Science*, pages 370–375, Berlin, Heidelberg, 2005. Springer
- R. Alberich, G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente. An algebraic metric for phylogenetic trees. *Applied Mathematics Letters*, 22(9):1320–1324, 2009

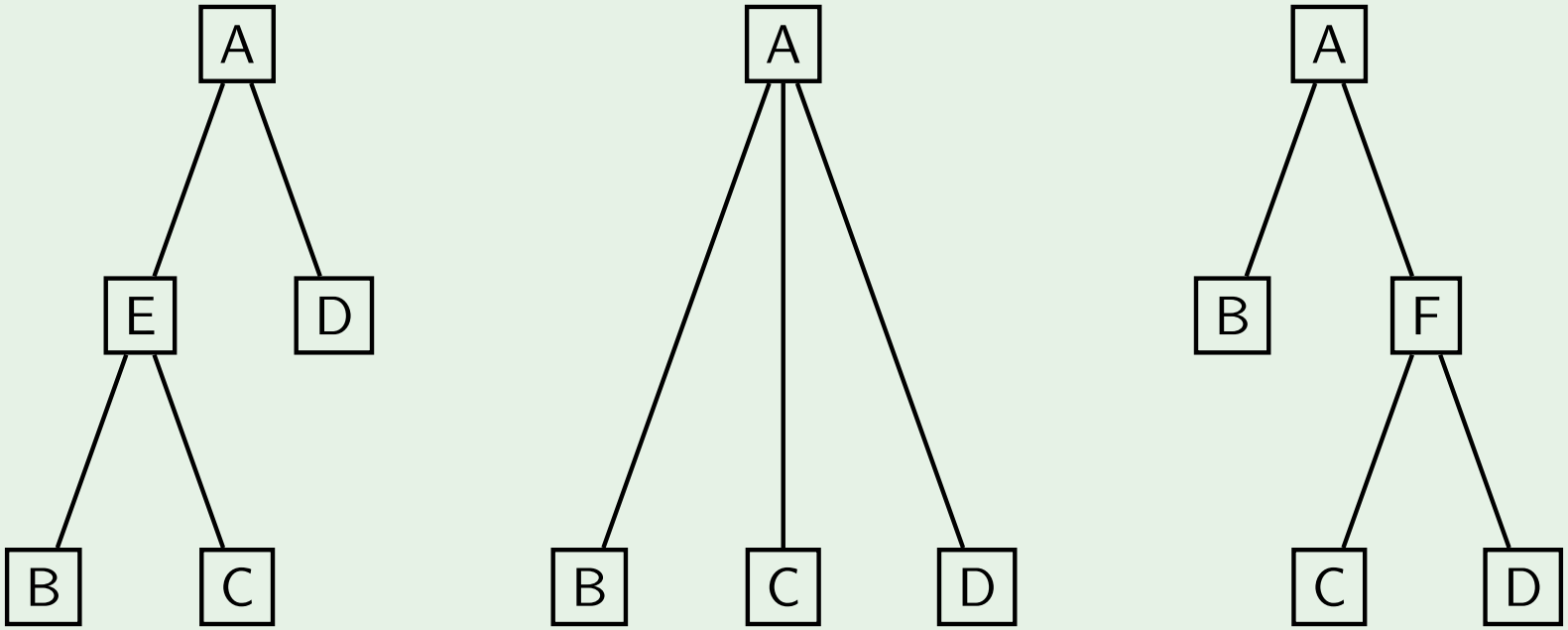
- The edit distance between two strings is the smallest number of insertions, deletions, and substitutions needed to transform one string into the other.
- An alignment of two strings is an arrangement of the two strings as rows of a matrix, with additional gaps (dashes) between the elements to make some or all of the remaining (aligned) columns contain identical elements but with no column gapped in both strings.

Example

```
-GCTTCCGGCTCGTATAATGTGTGG
 |||||*|*||    |||||* |
TGCTTCTGACT---ATAATA-G---
```

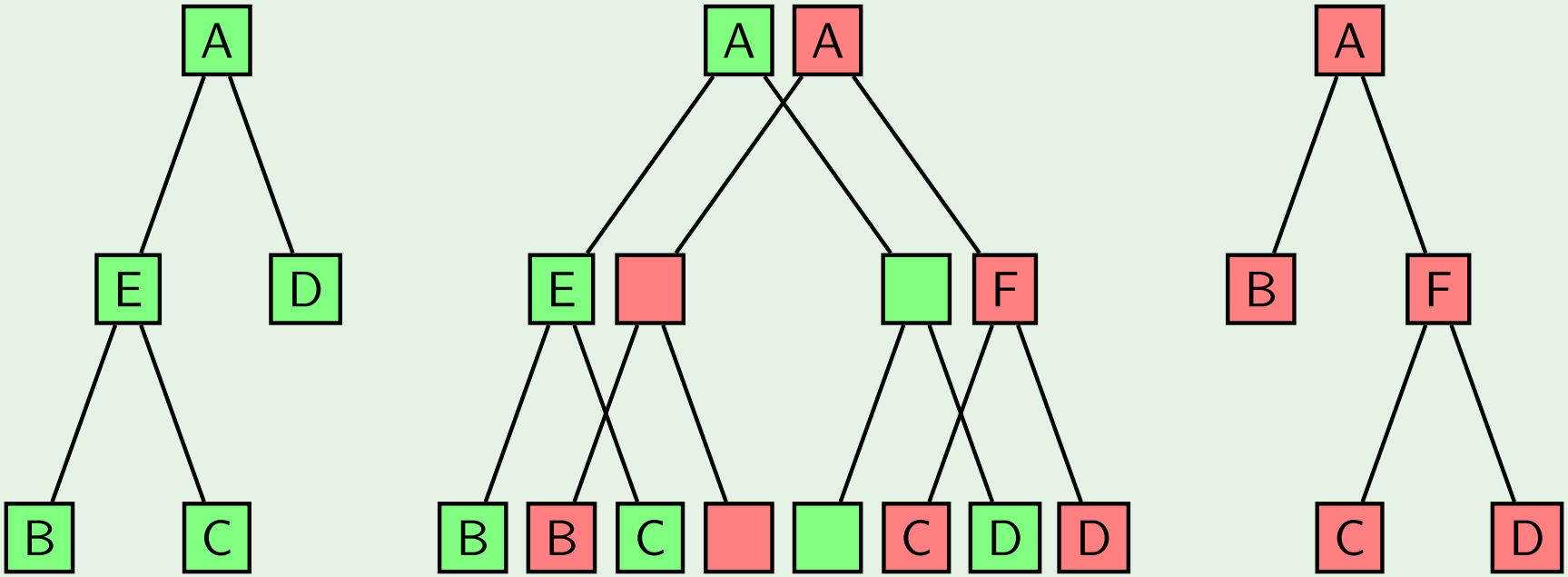
- The edit distance between two trees is the smallest number of insertions, deletions, and substitutions needed to transform one tree into the other.

Example



- An alignment of two trees is an arrangement of the trees with space labeled nodes inserted such that their structures coincide.

Example



- An alignment of trees is a restricted form of tree edit distance in which all the insertions precede all the deletions.
- With insertion cost 1, deletion cost 1, identical substitution cost 0, and non-identical substitution cost 2, an optimal tree edit yields a largest common subtree and an optimal alignment yields a smallest common supertree.

- T. Jiang, L. Wang, and K. Zhang. Alignment of trees—an alternative to tree edit. *Theoretical Computer Science*, 143(1):137–148, 1995

- A. Lozano, R. Pinter, O. Rokhlenko, G. Valiente, and M. Ziv-Ukelson. Seeded tree alignment and planar tanglegram layout. In *Proc. 7th Workshop on Algorithms in Bioinformatics*, volume 4645 of *Lecture Notes in Bioinformatics*, pages 98–110. Springer, 2007
- A. Lozano, R. Pinter, O. Rokhlenko, G. Valiente, and M. Ziv-Ukelson. Seeded tree alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(4):503–513, 2008