BIOINFORMATICS AND STATISTICAL GENETICS

GABRIEL VALIENTE

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

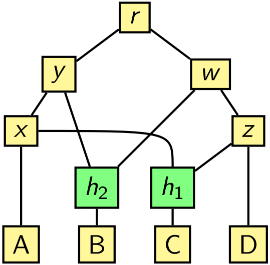
2023–2024

Agreement of phylogenetic networks

- Path multiplicity distance
- Tripartition distance
- Nodal distance
- Triplets distance
- Edit distance
- Alignment of phylogenetic networks

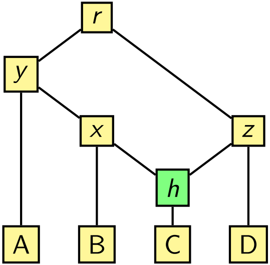
- The similarities and differences between two phylogenetic networks can be assessed by computing a distance measure between the two networks.
- The path multiplicity distance is based on the number of different paths from the internal nodes to each of the terminal nodes of the networks.
- While there is one (trivial) path from a terminal node to itself and no path to any other terminal node, the numbers of paths from the internal nodes to the terminal nodes reveal similarities and differences between two phylogenetic networks.

- Path multiplicity vectors for a fully resolved tree-child phylogenetic network, with the nodes sorted by height.



node	height	vector
A	0	$(1,0,0,0)$
B	0	$(0,1,0,0)$
C	0	$(0,0,1,0)$
D	0	$(0,0,0,1)$
h_1	1	$(0,0,1,0)$
h_2	1	$(0,1,0,0)$
x	2	$(1,0,1,0)$
z	2	$(0,0,1,1)$
w	3	$(0,1,1,1)$
y	3	$(1,1,1,0)$
r	4	$(1,2,2,1)$

- Path multiplicity vectors for another fully resolved tree-child phylogenetic network, with the nodes sorted by height.



node	height	vector
A	0	$(1, 0, 0, 0)$
B	0	$(0, 1, 0, 0)$
C	0	$(0, 0, 1, 0)$
D	0	$(0, 0, 0, 1)$
h	1	$(0, 0, 1, 0)$
x	2	$(0, 1, 1, 0)$
z	2	$(0, 0, 1, 1)$
y	3	$(1, 1, 1, 0)$
r	4	$(1, 1, 2, 1)$

- Their path multiplicity distance is 6.

node	height	vector
A	0	$(1,0,0,0)$
B	0	$(0,1,0,0)$
C	0	$(0,0,1,0)$
D	0	$(0,0,0,1)$
h_1	1	$(0,0,1,0)$
h_2	1	$(0,1,0,0)$
x	2	$(1,0,1,0)$
z	2	$(0,0,1,1)$
w	3	$(0,1,1,1)$
y	3	$(1,1,1,0)$
r	4	$(1,2,2,1)$

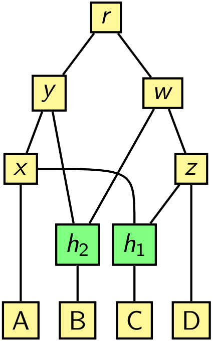
node	height	vector
A	0	$(1,0,0,0)$
B	0	$(0,1,0,0)$
C	0	$(0,0,1,0)$
D	0	$(0,0,0,1)$
h	1	$(0,0,1,0)$
x	2	$(0,1,1,0)$
z	2	$(0,0,1,1)$
y	3	$(1,1,1,0)$
r	4	$(1,1,2,1)$

- Notice that the path multiplicity vector $(0,1,0,0)$ occurs twice in the first network but only once in the second network and, thus, it contributes $|2-1|=1$ to the symmetric difference of the multisets of path multiplicity vectors.

- The path multiplicity distance between two phylogenetic networks labeled over the same taxa is defined as the size of the symmetric difference of their multisets of path multiplicity vectors, that is, the number of path multiplicity vectors in which the two phylogenetic networks differ.
- The symmetric difference applies to multisets rather than to sets, because path multiplicity vectors in a phylogenetic network are not necessarily unique.
- For instance, in a fully resolved phylogenetic network, a hybrid node and its single child share the same path multiplicity vector.

- Since the paths from an internal node to the terminal nodes of a phylogenetic network are the paths from the children of the internal node to the terminal nodes, the vector of path multiplicities associated with each node of a phylogenetic network can be computed by performing a bottom-up traversal, from the terminal nodes up to the root of the network, adding the path multiplicity vectors of the children to obtain the path multiplicity vector of the parent node.
- The path multiplicity vector of the i-th terminal node has 1 in the i-th position and 0 everywhere else.
- The path multiplicity vector $\mu(v)$ of each node v in a phylogenetic network N is computed during a bottom-up traversal of N, with the help of an (initially empty) queue Q of nodes.
- The path multiplicity vector $\mu(v)$ of each child v of an internal node u is added in turn to the (initially all-zero) path multiplicity vector $\mu(u)$ of the parent node u.

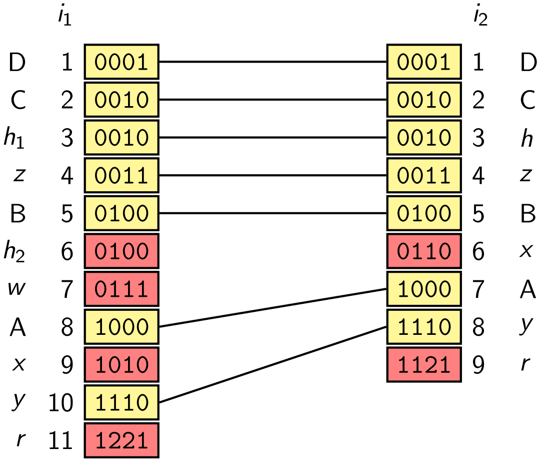
```
procedure path_multiplicity($N, \mu$)
    for all nodes $v$ of $N$ do
        $\mu(v) \leftarrow (0, 0, \ldots, 0)$
        if $v$ is a terminal node then
            $i \leftarrow$ rank of $v$ in the terminal nodes of $N$
            $\mu(v)[i] \leftarrow 1$
            $enqueue(Q, v)$
    while $Q$ is not empty do
        $v \leftarrow dequeue(Q)$
        mark node $v$ as visited
        for all parents $u$ of node $v$ do
            $\mu(u) \leftarrow \mu(u) + \mu(v)$
            if all children of $u$ are marked visited then
                $enqueue(Q, u)$
```



A	0	$\mu(x) \leftarrow (0,0,0,0) + (1,0,0,0) = (1,0,0,0)$
B	0	$\mu(h_2) \leftarrow (0,0,0,0) + (0,1,0,0) = (0,1,0,0)$
C	0	$\mu(h_1) \leftarrow (0,0,0,0) + (0,0,1,0) = (0,0,1,0)$
D	0	$\mu(z) \leftarrow (0,0,0,0) + (0,0,0,1) = (0,0,0,1)$
h_1	1	$\mu(x) \leftarrow (1,0,0,0) + (0,0,1,0) = (1,0,1,0)$ $\mu(z) \leftarrow (0,0,0,1) + (0,0,1,0) = (0,0,1,1)$
h_2	1	$\mu(w) \leftarrow (0,0,0,0) + (0,1,0,0) = (0,1,0,0)$ $\mu(y) \leftarrow (0,0,0,0) + (0,1,0,0) = (0,1,0,0)$
x	2	$\mu(y) \leftarrow (0,1,0,0) + (1,0,1,0) = (1,1,1,0)$
z	2	$\mu(w) \leftarrow (0,1,0,0) + (0,0,1,1) = (0,1,1,1)$
w	3	$\mu(r) \leftarrow (0,0,0,0) + (0,1,1,1) = (0,1,1,1)$
y	3	$\mu(r) \leftarrow (0,1,1,1) + (1,1,1,0) = (1,2,2,1)$

- The path multiplicity distance between two phylogenetic networks can be computed by counting the number of path multiplicity vectors shared by the two networks, during a simultaneous traversal of the sorted path multiplicity vectors of the two networks.
- The matrices of path multiplicities are sorted by rows and then the simultaneous traversal is performed by advancing the row index to the path multiplicity matrix of the first network, the second network, or both, depending on the indexed path multiplicity vector of the first network being less than, greater than, or equal to the indexed path multiplicity vector of the second network.
- In the latter case, the number of common path multiplicity vectors is increased by one.
- Then the path multiplicity distance is the number of nodes in the two networks minus twice the number of path multiplicity vectors shared by the two networks.

```
**function** path_multiplicity_distance($N_1, N_2$)
    *path_multiplicity*($N_1, \mu_1$)
    *path_multiplicity*($N_2, \mu_2$)
    sort $\mu_1$ and $\mu_2$
    $n_1, n_2 \leftarrow$ number of nodes of $N_1, N_2$
    $i_1 \leftarrow i_2 \leftarrow 1$
    $c \leftarrow 0$
    **while** $i_1 \leqslant n_1$ **and** $i_2 \leqslant n_2$ **do**
        **if** $\mu_1[i_1] < \mu_2[i_2]$ **then**
            $i_1 \leftarrow i_1 + 1$
        **else if** $\mu_1[i_1] > \mu_2[i_2]$ **then**
            $i_2 \leftarrow i_2 + 1$
        **else**
            $i_1 \leftarrow i_1 + 1$; $i_2 \leftarrow i_2 + 1$; $c \leftarrow c + 1$
    **return** $n_1 + n_2 - 2 \cdot c$
```



	i_1	
D	1	0001
C	2	0010
h_1	3	0010
z	4	0011
B	5	0100
h_2	6	0100
w	7	0111
A	8	1000
x	9	1010
y	10	1110
r	11	1221

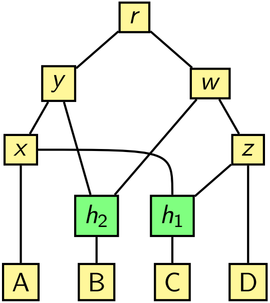
	i_2	
0001	1	D
0010	2	C
0010	3	h
0011	4	z
0100	5	B
0110	6	x
1000	7	A
1110	8	y
1121	9	r

- There are 7 path multiplicity vectors shared by the two networks and, thus, their path multiplicity distance is $11 + 9 - 2 \cdot 7 = 6$.

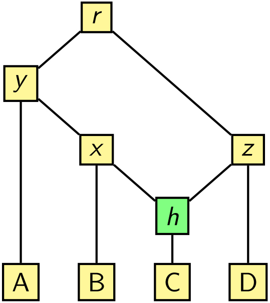
- The path multiplicity distance is a metric on the space of all tree-child phylogenetic networks, as well as on the space of all binary tree-sibling time-consistent phylogenetic networks.
- The path multiplicity distance generalizes the partition distance between rooted phylogenetic trees.
- The path multiplicity distance can be computed in $O(n^2)$ time.

- G. Cardona, F. Rosselló, and G. Valiente. Comparison of tree-child phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):552–569, 2009

- The tripartition distance is based on the partition of the taxa into strict descendants, non-strict descendants, and non-descendants induced by each node in the two phylogenetic networks under comparison.
- While terminal nodes are their only (strict) descendants, the tripartitions induced by the internal nodes reveal similarities and differences between two phylogenetic networks.



node	height	tripartition
A	0	(A, C, C, C)
B	0	(C, A, C, C)
C	0	(C, C, A, C)
D	0	(C, C, C, A)
h_1	1	(C, C, A, C)
h_2	1	(C, A, C, C)
x	2	(A, C, B, C)
z	2	(C, C, B, A)
w	3	(C, B, B, A)
y	3	(A, B, B, C)
r	4	(A, A, A, A)



node	height	tripartition
A	0	(A, C, C, C)
B	0	(C, A, C, C)
C	0	(C, C, A, C)
D	0	(C, C, C, A)
h	1	(C, C, A, C)
x	2	(C, A, B, C)
z	2	(C, C, B, A)
y	3	(A, A, B, C)
r	4	(A, A, A, A)

- The two phylogenetic networks differ in the tripartition vectors (A, A, B, C), (A, B, B, C), (A, C, B, C), (C, A, B, C), (C, A, C, C), (C, B, B, A) and, thus, their tripartition distance is 6.
- The tripartition distance between two phylogenetic networks labeled over the same taxa is defined as the size of the symmetric difference of their multisets of tripartition vectors, that is, the number of tripartition vectors in which the two phylogenetic networks differ.
- The symmetric difference applies to multisets rather than to sets, because tripartition vectors in a phylogenetic network are not necessarily unique.
- For instance, a hybrid node and its single child share the same tripartition vector in a fully resolved network.

- The multiset of tripartition vectors of a phylogenetic network can be obtained by testing for each node of the network if it is a strict ancestor, a non-strict ancestor, or not an ancestor at all of each of the terminal nodes in turn.

```
procedure tripartition($N, \theta$)
    for all nodes $v$ of $N$ do
        for all terminal nodes $w$ of $N$ do
            if $w$ is reachable from $v$ in $N$ then
                if $strict\_ancestor(N, v, w)$ then
                    $\theta[v][w] \leftarrow A$
                else
                    $\theta[v][w] \leftarrow B$
            else
                $\theta[v][w] \leftarrow C$
```

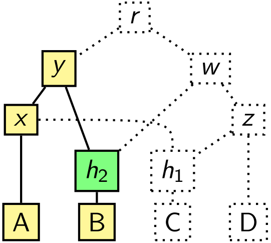
- As in the case of the path multiplicity distance, the tripartition distance between two phylogenetic networks can be computed by counting the number of tripartition vectors shared by the two networks during a simultaneous traversal of the sorted tripartition vectors of the two networks.
- The matrices of tripartitions are sorted by rows and then the simultaneous traversal is performed by advancing the row index to the tripartitions matrix of the first network, the second network, or both, depending on the indexed tripartition vector of the first network being less than, greater than, or equal to the indexed tripartition vector of the second network.
- In the latter case, the number of common tripartition vectors is increased by one.
- Then the tripartition distance is the number of nodes in the two networks minus twice the number of tripartition vectors shared by the two networks.

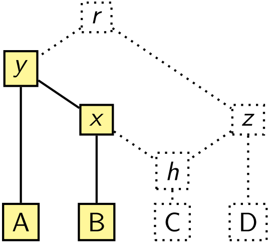
```
**function** tripartition_distance($N_1, N_2$)
    *tripartition*($N_1, \theta_1$)
    *tripartition*($N_2, \theta_2$)
    sort $\theta_1$ and $\theta_2$
    $n_1, n_2 \leftarrow$ number of nodes of $N_1, N_2$
    $i_1 \leftarrow i_2 \leftarrow 1$
    $c \leftarrow 0$
    **while** $i_1 \leqslant n_1$ **and** $i_2 \leqslant n_2$ **do**
        **if** $\theta_1[i_1] < \theta_2[i_2]$ **then**
            $i_1 \leftarrow i_1 + 1$
        **else if** $\theta_1[i_1] > \theta_2[i_2]$ **then**
            $i_2 \leftarrow i_2 + 1$
        **else**
            $i_1 \leftarrow i_1 + 1$; $i_2 \leftarrow i_2 + 1$; $c \leftarrow c + 1$
    **return** $n_1 + n_2 - 2 \cdot c$
```

- The tripartition distance is a metric on the space of all tree-child time-consistent phylogenetic networks.
- The tripartition distance generalizes the partition distance between rooted phylogenetic trees.
- The tripartition distance can be computed in $O(n^2)$ time.

- G. Cardona, F. Rosselló, and G. Valiente. Tripartitions do not always discriminate phylogenetic networks. *Mathematical Biosciences*, 211(2):356–370, 2008

- The nodal distance is based on the shortest paths between terminal nodes in the two phylogenetic networks under comparison.
- The matrices of distances between each pair of terminal nodes and their LCSA in the two networks reveal similarities and differences between two phylogenetic networks.





- The shortest path from the LCSA of A and B to B has length 2 in the two networks, while the shortest path to A has length 2 in the first network and length 1 in the second network.
- Absolute differences between the corresponding shortest path length matrices.

$$\left| \begin{pmatrix} 0 & 2 & 1 & 3 \\ 2 & 0 & 3 & 2 \\ 2 & 4 & 0 & 2 \\ 3 & 2 & 1 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 1 & 1 & 2 \\ 2 & 0 & 1 & 3 \\ 3 & 2 & 0 & 2 \\ 2 & 2 & 1 & 0 \end{pmatrix} \right| = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 2 & 1 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

- The nodal distance between the two phylogenetic networks is thus 9.

- The nodal distance between two phylogenetic networks labeled over the same taxa is defined as the sum of the absolute differences of distance between each pair of terminal nodes and their LCSA in the two networks.
- Therefore, the nodal distance can be obtained by computing the distance between each pair of terminal nodes and their LCSA in each of the two networks and then computing the absolute difference between the two matrices of nodal distances.
- The LCSA of each pair of terminal nodes in a network is computed only once, in order to obtain the two nodal distances between them.

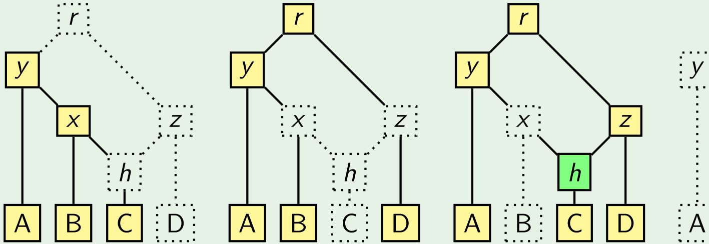
```
**function** nodal distance($N_1, N_2$)
    $L \leftarrow$ terminal node labels in $N_1$ and $N_2$
    $n \leftarrow length(L)$
    $d \leftarrow 0$
    **for** $i \leftarrow 1, \ldots, n$ **do**
        $i_1, i_2 \leftarrow$ terminal node of $N_1, N_2$ labeled $L[i]$
        **for** $j \leftarrow i + 1, \ldots, n$ **do**
            $j_1, j_2 \leftarrow$ terminal node of $N_1, N_2$ labeled $L[j]$
            $\ell_1 \leftarrow LCSA(N_1, i_1, j_1)$
            $\ell_2 \leftarrow LCSA(N_2, i_2, j_2)$
            $d_1 \leftarrow distance(N_1, \ell_1, i_1)$
            $d_2 \leftarrow distance(N_2, \ell_2, i_2)$
            $d \leftarrow d + |d_1 - d_2|$
            $d_1 \leftarrow distance(N_1, \ell_1, j_1)$
            $d_2 \leftarrow distance(N_2, \ell_2, j_2)$
            $d \leftarrow d + |d_1 - d_2|$
    **return** $d$
```

- The nodal distance is a metric on the space of all tree-child time-consistent phylogenetic networks.
- The nodal distance generalizes the nodal distance between rooted phylogenetic trees.
- The nodal distance can be computed in $O(n^4)$ time.

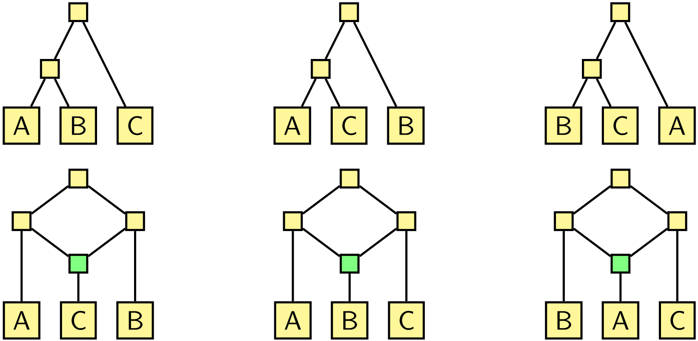
- G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente. Metrics for phylogenetic networks II: Nodal and triplets metrics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(3):454–469, 2009

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

Example



- A rooted phylogenetic network can be reconstructed (although not in a unique way) from the set of all its triplet topologies.
- Given a triplet of terminal nodes, there are only six possible induced subgraphs if the phylogenetic network is binary and time consistent.



- However, if the phylogenetic network is not time consistent, there is an infinite number of possible subgraphs induced by a triplet of terminal nodes, even if the phylogenetic network is binary.

```
function triplet($N, i, j, k$)
    $ij \leftarrow LCSA(N, i, j)$; $ik \leftarrow LCSA(N, i, k)$; $jk \leftarrow LCSA(N, j, k)$
    if $ij$ is an ancestor of $ik$ in $N$ then
        if $ik$ is an ancestor of $ij$ in $N$ then
            return $((j, k), i)$;
        else
            if $jk$ is an ancestor of $ij$ in $N$ then
                return $((i, k), j)$;
            else
                return $((i, (k)\#H1), (\#H1, j))$;
    else
        if $ik$ is an ancestor of $ij$ in $N$ then
            if $jk$ is an ancestor of $ij$ in $N$ then
                return $((i, j), k)$;
            else
                return $((i, (j)\#H1), (\#H1, k))$;
        else
            return $((j, (i)\#H1), (\#H1, k))$;
```

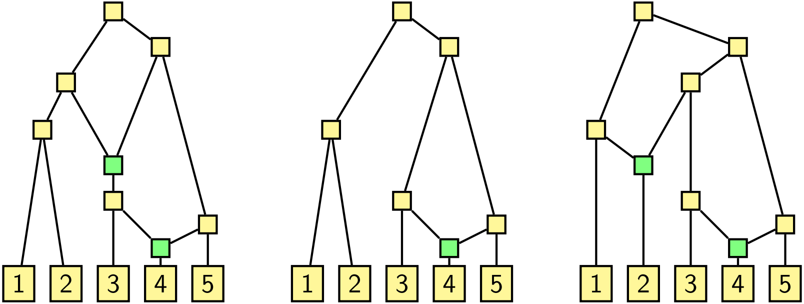
- The triplets distance between two phylogenetic networks can be computed by first obtaining the subtree induced by the LCSA of each set of three terminal nodes in each of the networks and then counting the number of induced subtrees in which the two networks differ.

```
**function** triplets_distance($N_1, N_2$)
    $L \leftarrow$ terminal node labels in $N_1$ and $N_2$
    $n \leftarrow length(L)$
    $d \leftarrow 0$
    **for** $i \leftarrow 1, \dots, n$ **do**
        **for** $j \leftarrow i + 1, \dots, n$ **do**
            **for** $k \leftarrow j + 1, \dots, n$ **do**
                $R_1 \leftarrow triplet(N_1, L[i], L[j], L[k])$
                $R_2 \leftarrow triplet(N_2, L[i], L[j], L[k])$
                **if** $R_1 \neq R_2$ **then**
                    $d \leftarrow d + 2$
    **return** $d$
```

- The triplets distance is a metric on the space of all tree-child time-consistent phylogenetic networks.
- The triplets distance generalizes the triplets distance between rooted phylogenetic trees.
- The triplets distance can be computed in $O(n^5)$ time.

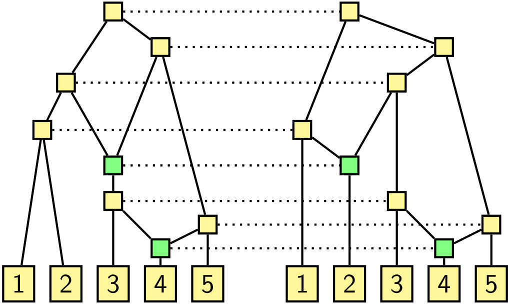
- G. Cardona, M. Llabrés, F. Rosselló, and G. Valiente. Metrics for phylogenetic networks II: Nodal and triplets metrics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(3):454–469, 2009
- J. Jansson, K. Mampentzidis, R. Rajaby, and W.-K. Sung. Computing the rooted triplet distance between phylogenetic networks. *Algorithmica*, 83(6):1786–1828, 2021

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



- M. Bordewicha, S. Linz, and C. Semple. Lost in space? Generalising subtree prune and regraft to spaces of phylogenetic networks. *Journal of Theoretical Biology*, 423(1):1–12, 2017
- J. Klawitter. The SNPR neighbourhood of tree-child networks. *Journal of Graph Algorithms and Applications*, 22(2):329–355, 2018

- An alignment of two tree-child phylogenetic networks is given by the solution to a bipartite matching problem, with edge weights based on the path multiplicity representation of the phylogenetic networks.
- An optimal alignment of two tree-child phylogenetic networks can be computed in $O(n^3)$ time.



- G. Cardona, F. Rosselló, and G. Valiente. Comparison of tree-child phylogenetic networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4):552–569, 2009

- 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