Definition

The notion of tree most often found in computational biology is that of a phylogenetic tree, that is, an either unrooted or rooted tree whose terminal nodes are labeled by taxa names.

In a classification tree, internal nodes may also be labeled by nested taxa.

These trees are often the product of clustering methods that attempt to reconstruct phylogenetic relationships from either distance data (estimates of the divergence between taxonomic units, in terms of number of mutations or some other quantitative measure of evolutionary change), such as the unweighted pair-group method with arithmetic mean (UPGMA) or the neighbor-joining (NJ) method, or DNA sequence data (making some assumption about the underlying model of DNA substitution), such as maximum parsimony, maximum likelihood, and Bayesian methods.

However, trees also arise as a mathematical model of RNA secondary structures without pseudo-knots.

Definition

The primary structure of ribonucleic acid (RNA) can be represented as a sequence over the alphabet of nucleotides: the purines A (adenine) and G (guanine), and the pyrimidines C (cytosine) and U (uracil).

Within the cell, RNA molecules do not retain such a linear form but, instead, fold back on themselves in space to form hydrogen bonds between short stretches of complementary nucleotide sequences: the most frequent ones are A–U and G–C, followed by the G–U bond.

The resulting secondary and tertiary structures are essential for RNA molecules to perform their biological roles.

Definition

The secondary structure of an RNA sequence can be seen as a set of nucleotide pairs $i - j$ with $i < j$ in which each nucleotide takes part in at most one pair, that is, such that $i = i'$ if and only if $j = j'$ for all nucleotide pairs $i - j$ and $i' - j'$.

This representation is also known as an arc-annotated sequence, and it is said to have no pseudo-knots when $i < i'$ if and only if $j > j'$ for all nucleotide pairs $i - j$ and $i' - j'$.

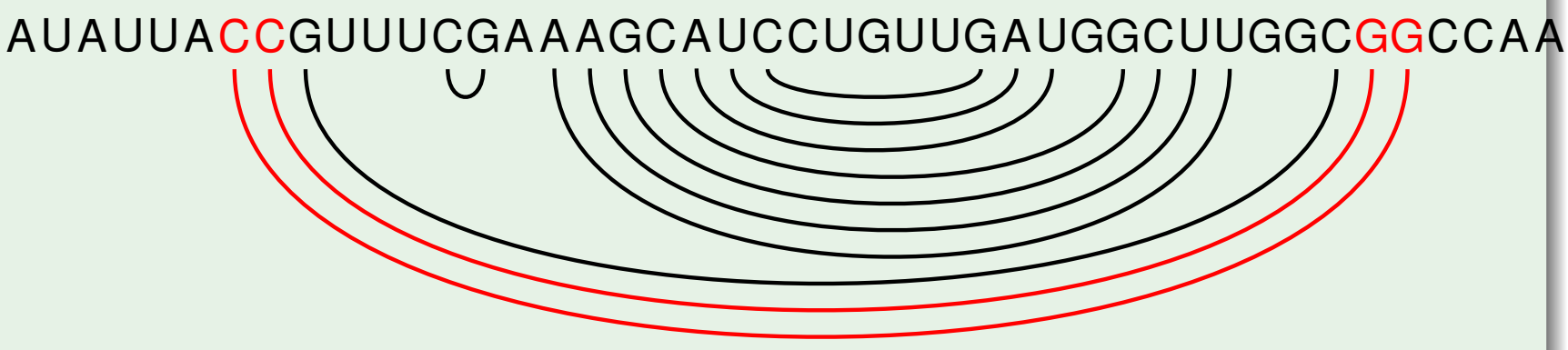
In an RNA structure without pseudo-knots, two nucleotide pairs $i - j$ and $i' - j'$ are said to be stacking if they are adjacent in the RNA sequence, that is, if $i' = i + 1$ and $j' = j - 1$.

Example

The secondary structure of the following RNA sequence, in 5′ to 3′ order,

```
AUAUUACCGUUUCGAAAGCAUCCUGUUGAUGGCUUGGCGGCCAA
```

corresponds to the arc-annotated sequence shown below.



There are no arc crossings in this arc-annotated sequence, because the secondary structure of this RNA sequence has no pseudo-knots. Among others, the nucleotide pairs C–G at sequence positions 7–40 and 8–39 are stacking.

Definition

Several types of RNA secondary structure elements can be distinguished:

Hairpin loops are unpaired stretches of nucleotides located within the two strands of the RNA molecule that end in an unpaired loop. They are defined by stacking nucleotide pairs $i-j, \ldots, i'-j'$ such that the inner nucleotides, that is, those between i' and j', are unpaired.

Internal loops are unpaired stretches of nucleotides located within the two strands of the RNA molecule. They are defined by two non-stacking nucleotide pairs $i-j$ and $i'-j'$ such that all the nucleotides between i and i' and also between j and j' are unpaired.

Bulge loops are unpaired stretches of nucleotides located within one strand of the RNA molecule. Bulges are thus a special class of internal loops which have no nucleotides either between i and i' or between j and j'. In a bulge, either $i' = i+1$ or $j' = j-1$.

Definition

Multiple bifurcation loops are also unpaired stretches of nucleotides located within the two strands of the RNA molecule, but they are defined by three or more non-stacking nucleotide pairs $i - j$, $i' - j'$, and $i'' - j''$ such that all the nucleotides between i and i' and also between j and j' and between i'' and j'' are unpaired.

These secondary structure elements are joined by helical stems of paired nucleotides within the RNA molecule.

Several methods are known for computing an RNA secondary structure from sequence data, and the predicted structures depend on either thermodynamic data or phylogenetic comparison, that is, the determination of those structural features that are conserved during evolution.

Example

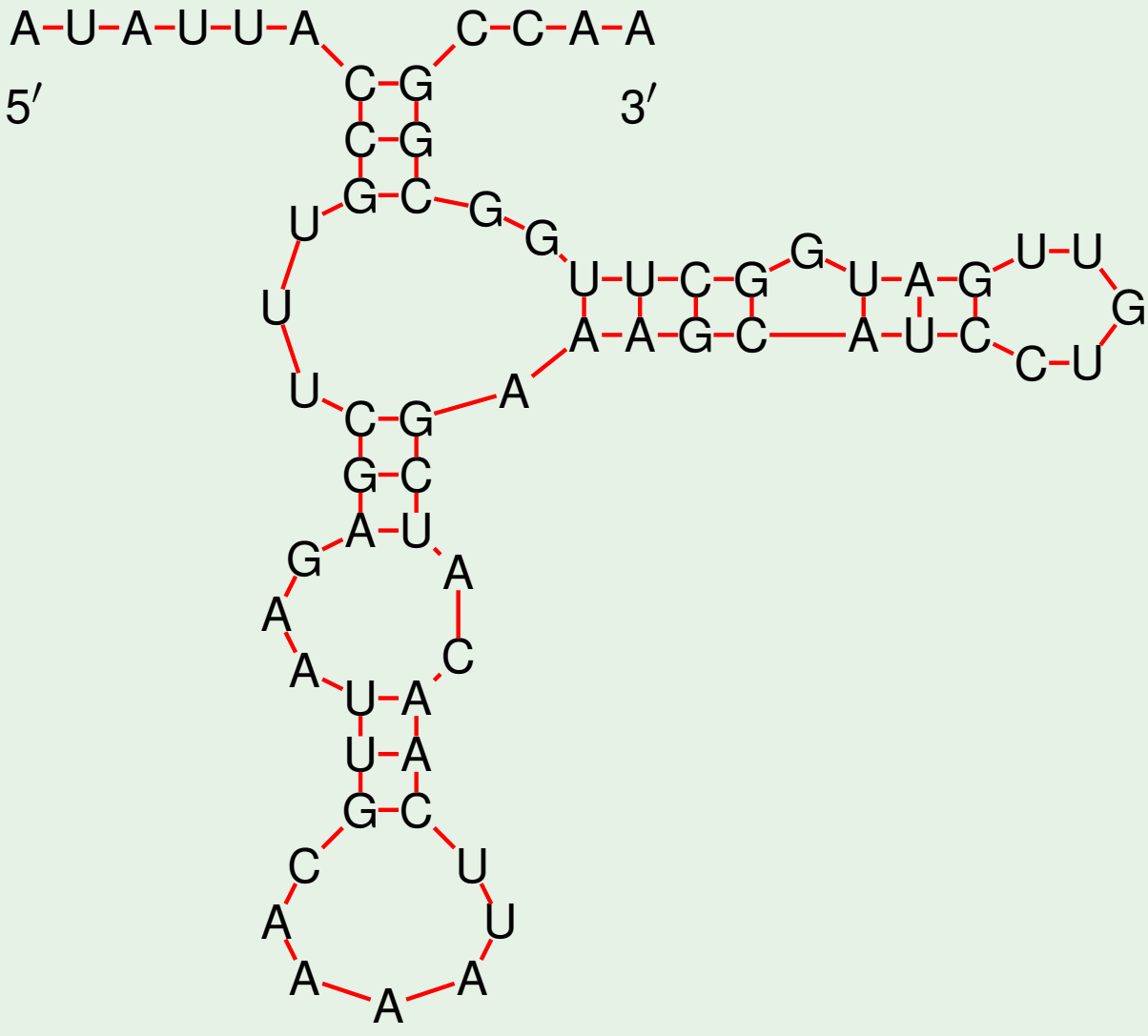
The following RNA sequence, in 5′ to 3′ order,

```
AUAUUACCGUUUCGAGAAUUGCAAAAUUCAACAUCGAAAGCAUCCUGUUGAUG
GCUUGGCGGCCAA
```

may fold into the predicted RNA secondary structure shown next.

There is a single-stranded region of 6 and 4 nucleotides followed by a helical stem of 3 paired nucleotides, a bifurcating loop of 6 nucleotides, and, along the vertical branch, a helical stem of 3 paired nucleotides, an internal loop of 5 nucleotides, another helical stem of 3 paired nucleotides, and a hairpin loop of 7 nucleotides, while along the horizontal branch there is a helical stem of 4 paired nucleotides, a small bulge loop of only 1 nucleotide, another helical stem of 3 paired nucleotides, and a hairpin loop of 5 nucleotides.

Example


A U A U U A
C C A A
5′
3′
C G
C G
G C
G
G
U
U
U
C G
G C
A U
G
A
A
C
A U
U A
G C
C
U
A
U
A
A
A
U U C G
G U A G
A A G C
A U C
C U
G
A

Definition

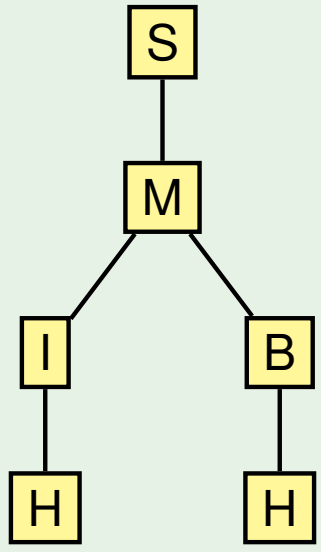
In an abstract model, RNA secondary structures without pseudo-knots can be represented by means of trees in which the ordering among sibling nodes corresponds to the $5'$ to $3'$ order of the RNA sequence.

There is a node in such a tree for each secondary structure element of the RNA molecule, labeled by H (hairpin loop), B (bulge), I (internal loop), M (multiple bifurcation loop), or S (single-stranded region).

Branches in the tree correspond to the helical stems that join secondary structure elements in the RNA molecule.

Example

The following ordered labeled tree is an abstract representation of the RNA secondary structure from the previous example.



Definition

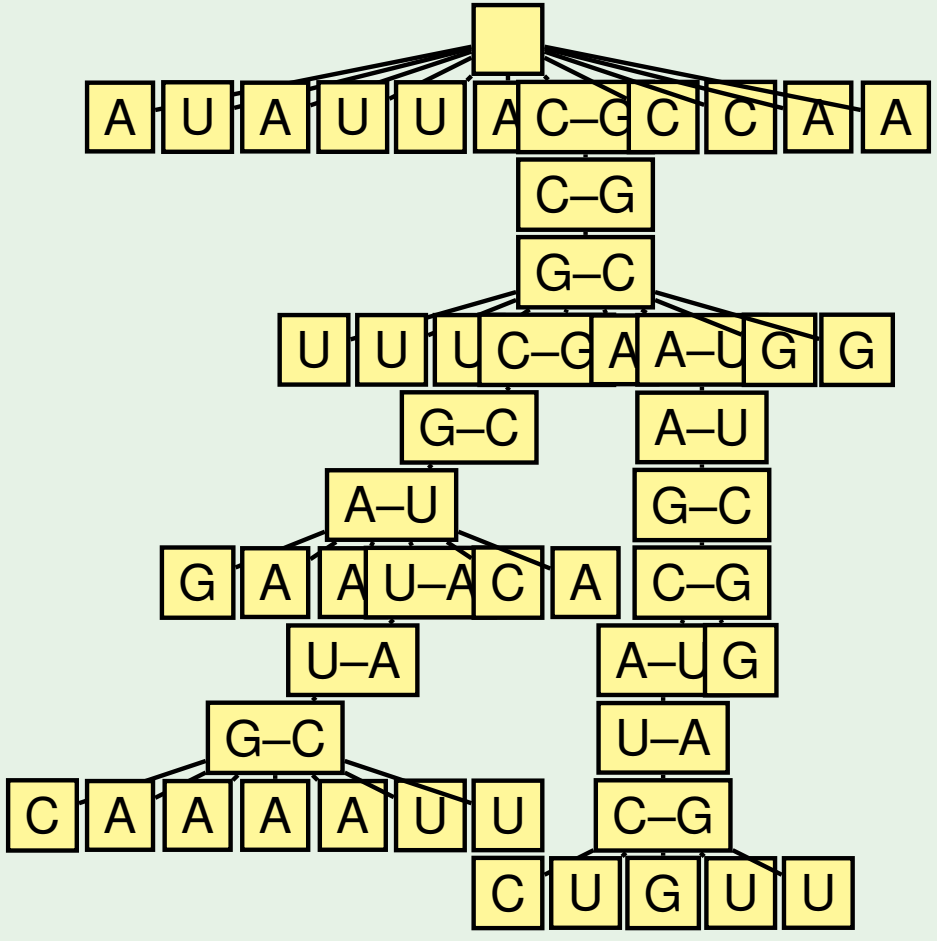
In a more detailed model, RNA secondary structures without pseudo-knots are also represented by means of trees in which the ordering among sibling nodes corresponds to the $5'$ to $3'$ order of the RNA sequence, but where there is a node for each unpaired nucleotide (the leaves) or paired nucleotide (the internal nodes of the tree).

Branches in the tree correspond to consecutive helical stems, and they also join the nucleotides of a secondary structure element with the previous secondary structure element along the RNA sequence.

The whole tree is rooted at a special, additional node.

Example

The following ordered labeled tree is a mode detailed representation of the RNA secondary structure from the previous examples.



Definition

The evolutionary relationships among a group of organisms are often illustrated by means of a phylogenetic tree, also called a cladogram or a dendrogram.

The nodes of a phylogenetic tree represent taxonomic units, which can be species or taxa, higher or nested taxa, populations, individuals, or genes.

The branches of a phylogenetic tree define the evolutionary relationships among the taxonomic units, in such a way that children nodes descend from their parent.

Such a pattern of ancestry and descent relationships is called the topology of the phylogenetic tree.

Definition

The root of a phylogenetic tree represents the most recent common ancestor of the taxonomic units at the leaves of the tree, and it is called fully resolved if every internal node in the tree has exactly two children.

In the case of ancient ancestry, though, such information is not always available.

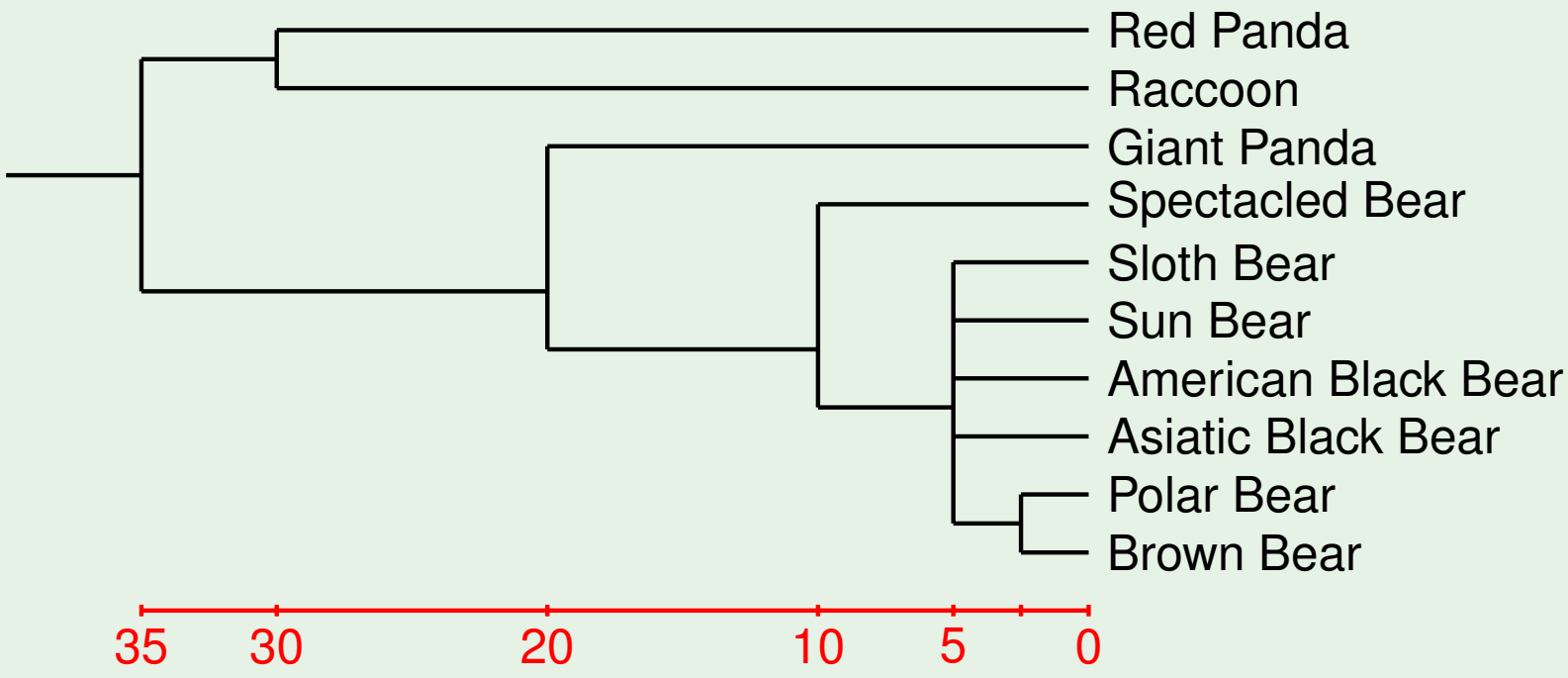
Moreover, most phylogenetic tree reconstruction methods yield unrooted trees.

In such a case, the phylogenetic tree can be rooted by choosing an outgroup and then placing the root between the outgroup and the node connecting it to the ingroup.

An outgroup is a taxonomic unit for which there is additional knowledge (such as taxonomic or paleontological information) about its divergence from the common ancestor prior to all the other taxonomic units, which then become the ingroup of the phylogenetic tree.

Example

The phylogenetic relationships among pandas, bears, and raccoons, assessed using mitochondrial DNA sequence evolution, indicate an early divergence (about 30 million years ago) of the red panda from the raccoon, whereas these species diverged from the outgroup of other carnivore families about 35 million years ago.

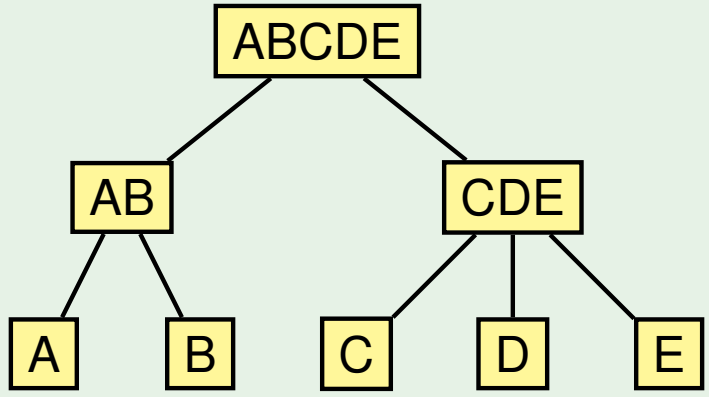


Definition

The Newick format is the de facto standard for representing phylogenetic trees, and it is quite convenient since it makes it possible to describe a whole tree in linear form in a unique way once the tree is drawn or the ordering among children nodes is fixed.

The Newick description of a tree is a string of nested parentheses annotated with taxa names and possibly also with branch lengths or bootstrap values (which measure how consistently the phylogenetic tree topology is supported by the underlying data).

Example




```
((A,B)AB,(C,D,E)CDE)ABCDE;
```

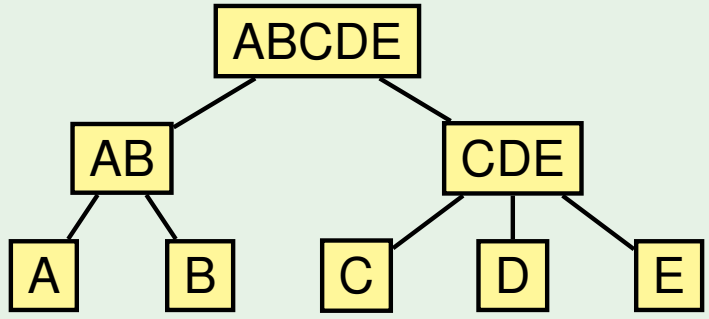
Algorithm

The Newick description of a given tree can be obtained by traversing the tree in postorder and writing down the name or label of the node when visiting a terminal (taxon) node, a left parenthesis (preceded by a comma unless the node is the first child of its parent) when visiting a non-terminal node for the first time, and a right parenthesis followed by the name or label of the node (if any) when visiting a non-terminal node for the second time, that is, after having visited all its descendants.

The name of a node is preceded by a comma unless it is the first child of its parent, and it is followed by a colon and the length (if any) of the branch from its parent.

The description of the tree is terminated with a semicolon.

Example

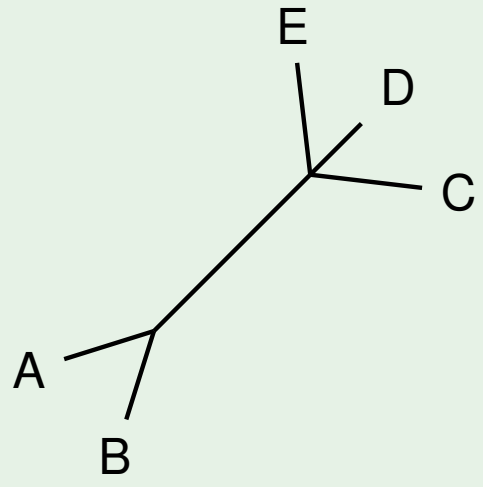


first visit non-terminal node ABCDE	`(`
first visit non-terminal node AB	`((`
visit terminal node A	`((A`
visit terminal node B	`((A,B`
second visit non-terminal node AB	`((A,B)AB`
first visit non-terminal node CDE	`((A,B)AB,(`
visit terminal node C	`((A,B)AB,(C`
visit terminal node D	`((A,B)AB,(C,D`
visit terminal node E	`((A,B)AB,(C,D,E`
second visit non-terminal node CDE	`((A,B)AB,(C,D,E)CDE`
second visit non-terminal node ABCDE	`((A,B)AB,(C,D,E)CDE)ABCDE;`

Definition

In the Newick representation of an unrooted phylogenetic tree, there are at least three siblings connected to some internal node.

Example

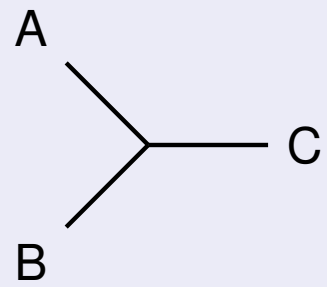



```
((A,B),C,D,E);
(A,B,(C,D,E));
```

Definition

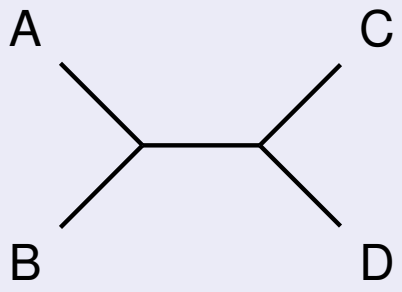
The number of possible phylogenetic trees also increases very rapidly with the number of terminal nodes, and for 10 terminal nodes there are already more than 2 million fully resolved unrooted phylogenetic trees and more than 34 million fully resolved rooted phylogenetic trees.

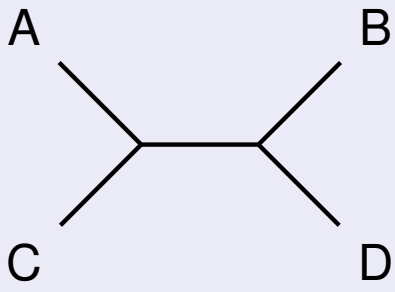
There is only 1 way to connect three labeled nodes A, B, C to make an unrooted phylogenetic tree.

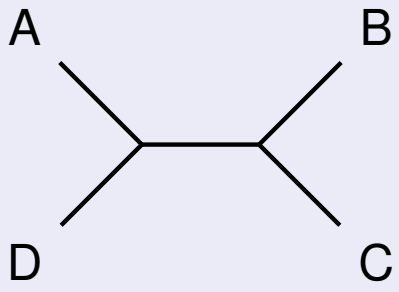


Definition

Four labeled nodes A, B, C, D can be connected in $1 \cdot 3 = 3$ different ways to make a fully resolved unrooted phylogenetic tree.







In general, a fully resolved unrooted phylogenetic tree with $n \geqslant 3$ terminal nodes has $n-2$ internal nodes and $2n-3$ branches, each of which can be split to accommodate a branch to a new terminal node.

Thus, the number $U(n)$ of fully resolved unrooted phylogenetic trees with $n \geqslant 3$ terminal nodes is $U(n) = U(n-1)(2(n-1)-3) = U(n-1)(2n-5) = \Pi_{i=3}^{n}(2i-5) = 1 \cdot 3 \cdot 5 \cdots (2n-5) = (2n-5)!!$

Double factorial

The double factorial notation

$$\begin{aligned}(2n)!! &= 2\cdot 4\cdot 6\cdots(2n-2)\cdot(2n) = 2^n n! \\ (2n+1)!! &= 1\cdot 3\cdot 5\cdots(2n-1)\cdot(2n+1) = \frac{(2n+1)!}{2^n n!}\end{aligned}$$

may be considered as a generalization of $n! = 1\cdot 2\cdot 3\cdots n$.

- B. E. Meserve. Double factorials.
  The American Mathematical Monthly, 55(7):425–426, 1948

Definition

On the other hand, each of the branches of a fully resolved unrooted phylogenetic tree can be split to accommodate a branch to the root and, thus, the number $T(n-1)$ of rooted phylogenetic trees with $n-1 \geqslant 3$ terminal nodes is the same as the number $U(n)$ of unrooted phylogenetic trees with $n \geqslant 3$ terminal nodes.

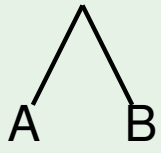
Thus, the number $T(n)$ of rooted phylogenetic trees with $n \geqslant 2$ terminal nodes is $T(n) = U(n+1) = (2(n+1)-5)!! = (2n-3)!!$

Algorithm

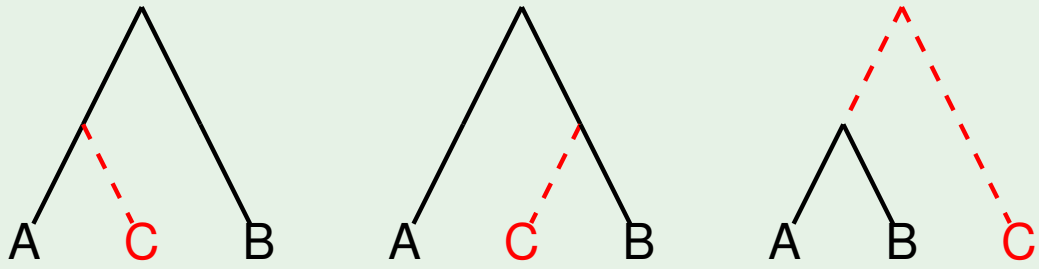
All the fully resolved rooted phylogenetic trees on $n \geqslant 2$ terminal nodes can be generated by taking each of the fully resolved rooted phylogenetic trees on $n-1$ terminal nodes in turn and then either splitting one of the branches to the n-th terminal node or joining it together with the n-th terminal node at a new root.

Example

The three fully resolved rooted phylogenetic trees on three terminal nodes can be generated by taking the only fully resolved rooted phylogenetic tree topology on two terminal nodes,



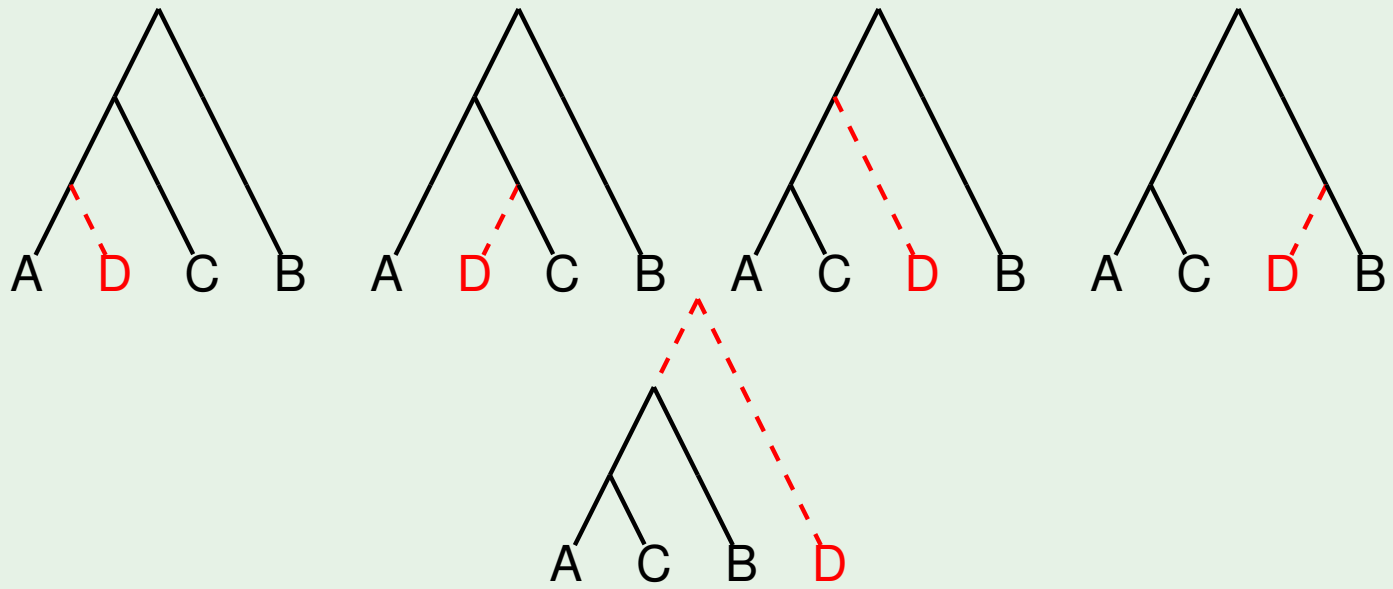
and either splitting one of the two branches to the new terminal node C or joining (A, B) with C at a new root.

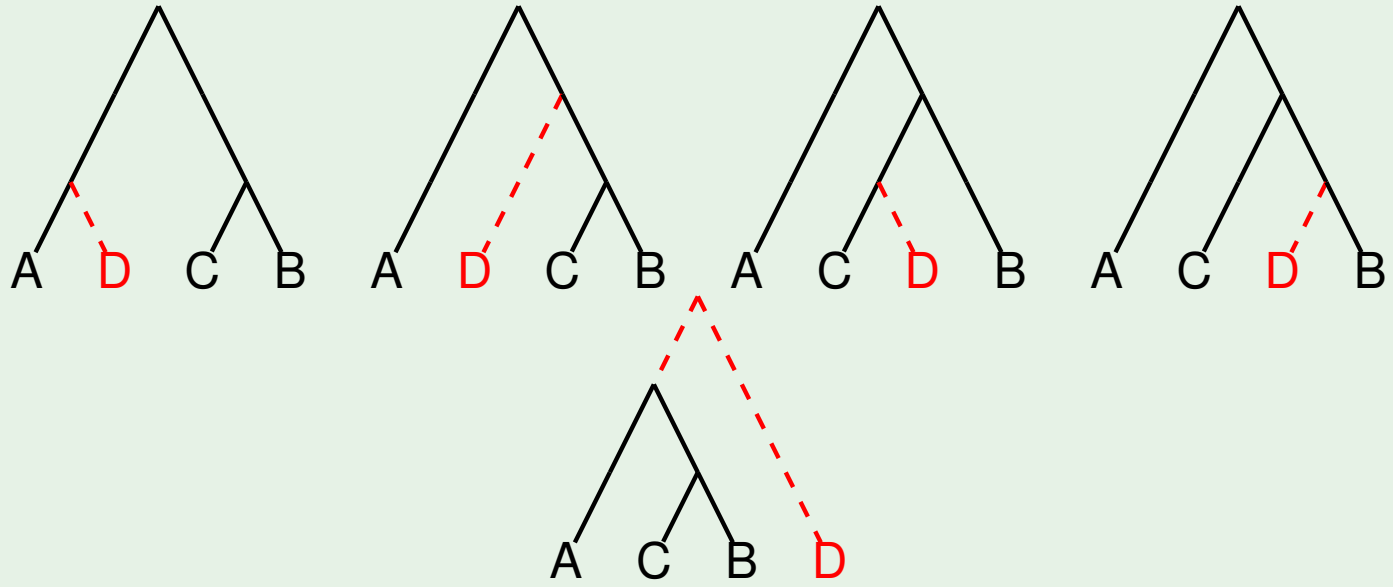


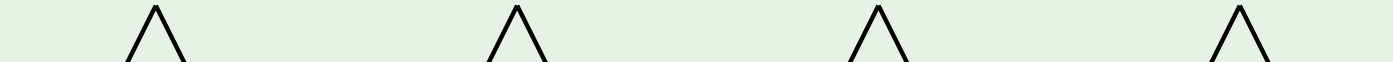
Example

The 15 fully resolved rooted phylogenetic trees on 4 terminal nodes can be generated by taking each of these three fully resolved rooted phylogenetic tree topologies on three terminal nodes and either splitting one of the four branches to the new terminal node D or joining the tree with D at a new root.

Example







Definition

There are also many ways in which trees can be represented in R and, as a matter of fact, many different R contributed packages implementing various types of trees are available for download from CRAN, the Comprehensive R Archive Network, at `http://cran.r-project.org/`. Among them, let us focus on the APE (Analysis of Phylogenetics and Evolution) tree representation, which is essentially a matrix-based representation of phylogenetic trees.

A phylogenetic tree in represented in the R package APE as a list of class `phylo` consisting of three elements: a numeric matrix `edge` with two columns and one row for each branch in the tree, a character vector `tip.label` with the labels of the internal nodes, and the number `Nnode` of internal nodes in the tree. In the edge matrix, all internal nodes appear in the first column at least twice, and values corresponding to terminal nodes appear only in the second column.

This representation is shared by rooted and unrooted phylogenetic trees. For the former, in a tree with n terminal nodes the root is numbered $n+1$, while for the latter, the numbering of the root is arbitrary. In the edge matrix, all nodes (except the root) appear exactly once in the second column, corresponding to the only branch from their parent.

Example

The phylogenetic tree with Newick string `((A,B)C);` has three terminal nodes numbered 1, 2, 3 and labeled A, B, C; two internal nodes numbered 4, 5; and branches 4–5, 5–1, 5–2, 4–3. The following R script computes the representation of such a phylogenetic tree, where the edge matrix is given by default in column order.

```
> library(ape)
> tree <- list(
    edge = matrix(c(4,5,5,4,5,1,2,3),4,2),
    tip.label = c("A","B","C"),
    Nnode = 2)
> class(tree) <- "phylo"
> tree

Phylogenetic tree with 3 tips and 2 internal nodes.

Tip labels:
[1] "A" "B" "C"

Rooted; no branch lengths.
```

Example

The same tree representation can be obtained by reading in a Newick string.

```
> tree <- read.tree(text="((A,B),C);")
> tree

Phylogenetic tree with 3 tips and 2 internal nodes.

Tip labels:
[1] "A" "B" "C"

Rooted; no branch lengths.
```

Example

A phylogenetic tree can also be reconstructed with the R package APE from distance data, using, for instance, the neighbor-joining method,

```
> mat <- matrix(scan("distances.mat"),n,n,byrow=T)
> tree <- nj(mat)
```

and it can be reconstructed from aligned DNA sequence data, using, for instance, the Kimura model of DNA substitution together with neighbor-joining.

```
> aln <- read.dna("sequences.aln")
> mat <- dist.dna(aln, model="K80")
> tree <- nj(mat)
```

Example

The branches of a phylogenetic tree can be obtained by accessing the `edge` matrix.

```
> tree$edge
     [,1] [,2]
[1,]    4    5
[2,]    5    1
[3,]    5    2
[4,]    4    3
```

The labels of the terminal nodes can also be obtained by accessing the `tip.label` character vector.

```
> tree$tip.label
[1] "A" "B" "C"
```

In the same way, the number of internal nodes can be obtained by accessing the `Nnode` variable.

```
> tree$Nnode
[1] 2
```

Example

Based on this matrix representation, it is rather easy to code various operations on phylogenetic trees and their nodes and branches; for instance, to obtain the number of terminal nodes,

```
> length(tree$tip.label)
[1] 3
```

to determine the number of branches,

```
> dim(tree$edge)[1]
[1] 4
```

Example

to find (the position, by column order, in the edge matrix of) a node in the tree,

```
> which(tree$edge == 1)
[1] 6
> which(tree$edge == 2)
[1] 7
> which(tree$edge == 3)
[1] 8
> which(tree$edge == 4)
[1] 1 4
> which(tree$edge == 5)
[1] 2 3 5
```

Example

to access the (branches from the) root of the tree,

```
> which(tree$edge[,1]==length(tree$tip.label)+1)
[1] 1 4
> tree$edge[
    which(tree$edge[,1]==length(tree$tip.label)+1),]
     [,1] [,2]
[1,]    4    5
[2,]    4    3
```

and to obtain the parent of a node in the tree.

```
> tree$edge[which(tree$edge[,2] == 1),1]
[1] 5
> tree$edge[which(tree$edge[,2] == 2),1]
[1] 5
> tree$edge[which(tree$edge[,2] == 3),1]
[1] 4
> tree$edge[which(tree$edge[,2] == 5),1]
[1] 4
```

Example

These operations can certainly be wrapped in a function.

```
> parent <- function (tree,x)
    tree$edge[which(tree$edge[,2] == x),1]
> parent(tree,1)
[1] 5
> parent(tree,2)
[1] 5
> parent(tree,3)
[1] 4
> parent(tree,5)
[1] 4
```

Example

Phylogenetic trees can also be displayed using R in a variety of ways, such as in Newick format,

```
> write.tree(tree)
[1] "((A,B),C);"
```

with the Newick string stored in a file,

```
> write.tree(tree, file = "tree.tre")
```

and drawn as a rectangular cladogram, with horizontal orientation and ancestral nodes centered over their descendants, in Encapsulated PostScript (EPS) format,

```
> postscript(file="cladogram.eps")
> plot(tree, type = "p")
> dev.off()
```

among several other display options for unrooted and rooted phylogenetic trees as well.

Combinatorial pattern matching is the search for exact or approximate occurrences of a given pattern within a given text.

When it comes to trees in computational biology, both the pattern and the text are trees and the pattern matching problem becomes one of finding the occurrences of a tree within another tree.

For instance, scanning an RNA secondary structure for the presence of a known pattern can help in finding conserved RNA motifs, and finding a phylogenetic tree within another phylogenetic tree can help in assessing their similarities and differences.

A related pattern matching problem that arises in the analysis of trees consists in finding simpler patterns, that is, paths within a given tree.

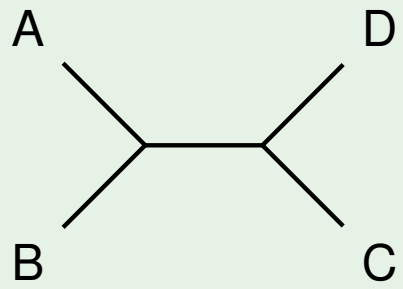
For instance, finding the path between two nodes of a tree is useful for computing distances in a tree and also for computing distances between two trees.

Definition

Any two nodes are connected by exactly one path in a tree, as long as no branch is to be traversed more than once in the path between the two nodes of the tree.

Example

In the following fully resolved unrooted phylogenetic tree, the path between terminal nodes A and D involves three branches.



The path A–D–B–C is not valid because the internal branch is traversed more than once: it is traversed three times along this path.

Definition

The path between any two given terminal nodes of a fully resolved rooted phylogenetic tree traverses the most recent common ancestor of the nodes in the tree.

Algorithm

The most recent common ancestor of two terminal nodes in a rooted phylogenetic tree can be found by obtaining first the lineages (paths to the root) of the two terminal nodes in the tree and then finding the first node in one of the lineages than also belongs to the other lineage.

The lineage of node i in a rooted phylogenetic tree T is stored in a list L and then the nodes in the lineage of node j are tested one after the other until a node is found that also belongs to the lineage of node i.

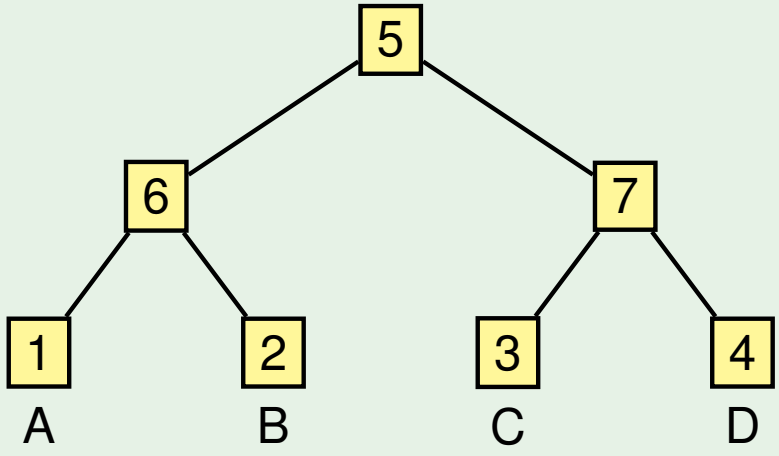
Algorithm

function mrca(T, i, j)
    $L \leftarrow \{i\}$
    $k \leftarrow i$
    while $T[k] \neq root(T)$ **do**
        $k \leftarrow parent(T, k)$
        $L \leftarrow L \cup \{k\}$
    $k \leftarrow j$
    while $T[k] \neq root(T)$ **do**
        if $k \in L$ **then**
            return k
        else
            $k \leftarrow parent(T, k)$

Example

In the following fully resolved rooted phylogenetic tree, the most recent common ancestor of each pair of terminal nodes is indicated in the table to the right.

The four terminal nodes labeled A through D are numbered 1 through 4, respectively, and the three internal nodes are numbered 5 through 7, for reference.



	A	B	C	D
A	1	6	5	5
B	6	2	5	5
C	5	5	3	7
D	5	5	7	4

Definition

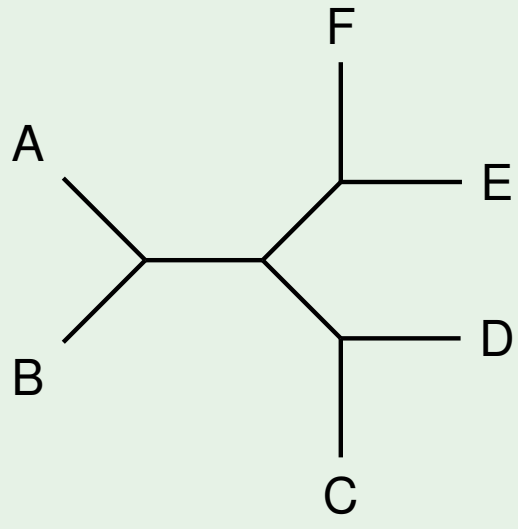
The distance between any two terminal nodes in an unrooted phylogenetic tree is the length of the path between the two terminal nodes in the tree.

The distance between any two terminal nodes in a rooted phylogenetic tree is the sum of the length of the paths between the two nodes and their most recent common ancestor in the tree.

In the case of phylogenetic trees with branch lengths, the distance between any two terminal nodes can be calculated as the sum of the length of the branches in the path between the two terminal nodes in the tree.

Example

The distance between each pair of terminal nodes in the following fully resolved unrooted phylogenetic tree is indicated in the table to the right.



	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

Definition

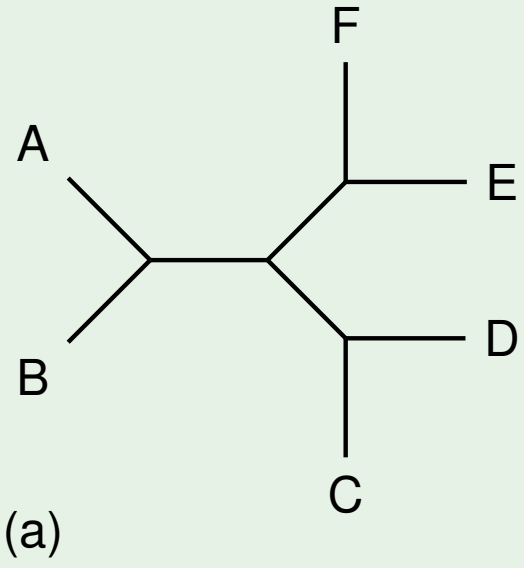
The similarities and differences between two unrooted phylogenetic trees can be assessed by computing a distance measure between the two trees.

The partition distance is based on the partition of the taxa induced by each internal branch in the two trees under comparison.

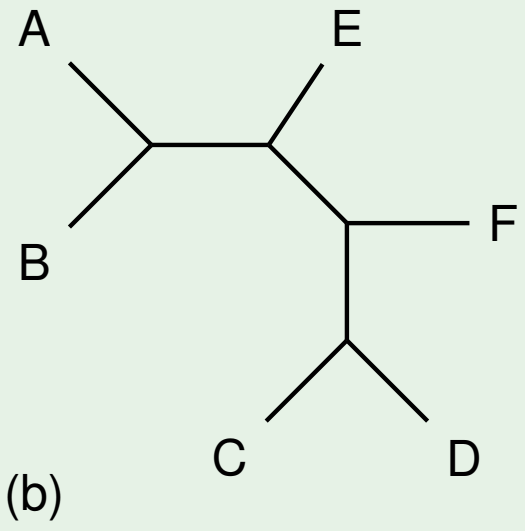
While cutting a tree along each of the the external branches partitions the set of taxa in a trivial way (with each partition consisting of a single taxon on one side and the remaining taxa on the other side), cutting along each of the internal branches reveals similarities and differences between the two trees.

Example

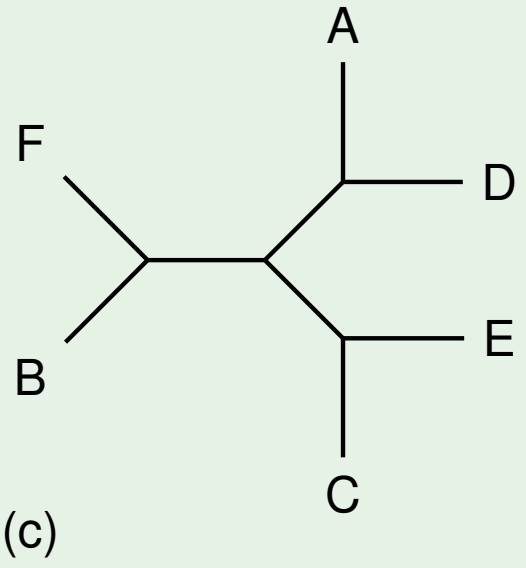
Each of the following fully resolved unrooted phylogenetic trees has six taxa and three internal branches and, thus, can be partitioned in three different ways by cutting one of the internal branches.



(a)



(b)



(c)

(a)	(A, B) and (C, D, E, F)	(A, B, C, D) and (E, F)	(A, B, E, F) and (C, D)
(b)	(A, B) and (C, D, E, F)	(A, B, E) and (C, D, F)	(A, B, E, F) and (C, D)
(c)	(A, B, D, F) and (C, E)	(A, C, D, E) and (B, F)	(A, D) and (B, C, E, F)

Definition

The partition distance between two unrooted phylogenetic trees is defined as the size of the symmetric difference of the partitions of the set of taxa obtained when cutting the trees along each of their internal branches, that is, the number of internal branches in the two trees resulting in different partitions of the taxa.

Example

In the previous example, there are two identical partitions between trees (a) and (b), while neither tree (a) nor tree (b) share any partition with tree (c).

There are two different partitions between trees (a) and (b), the partition (A, B, C, D) and (E, F) in tree (a) and the partition (A, B, E) and (C, D, F) in tree (b) and, thus, the partition distance between trees (a) and (b) is 2, while the partition distance between trees (a) and (c) and between trees (b) and (c) is 6, and, therefore, tree (a) is more similar to tree (b) than to tree (c).

Algorithm

The partition distance between two unrooted 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.

In the rooted representation of an unrooted tree, 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.

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

Algorithm

Once the partition of the taxa in each of the two phylogenetic trees induced by each of their internal branches is known, the partition distance can be computed by 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 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

Definition

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,

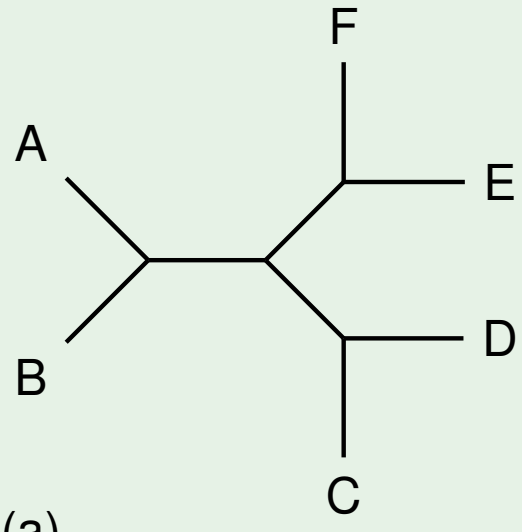
$$D(T) = (d_T(1,2), d_T(1,3), \ldots, d_T(1,n), d_T(2,3), \ldots, d_T(n-1,n)),$$

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,

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

Example

The distances between each two terminal nodes in the following fully resolved unrooted phylogenetic trees are given in the tables next to each of the trees.

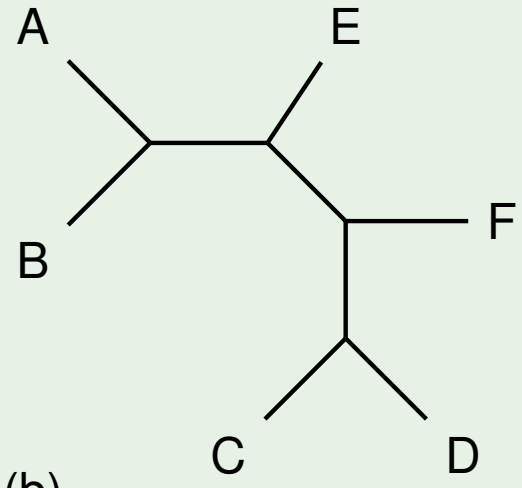


(a)

	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

The distances between each two terminal nodes in the following fully resolved unrooted phylogenetic trees are given in the tables next to each of the trees.

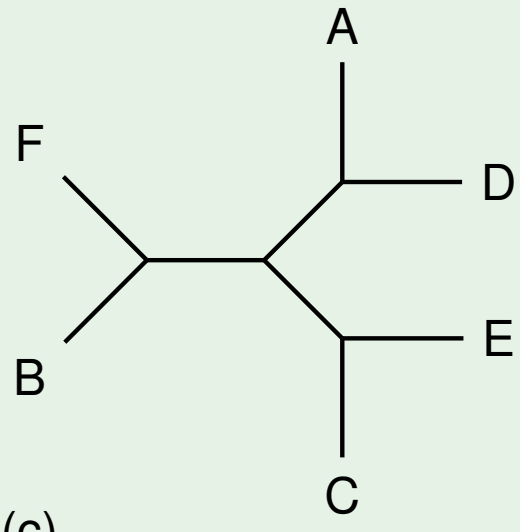


(b)

	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

The distances between each two terminal nodes in the following fully resolved unrooted phylogenetic trees are given in the tables next to each of the trees.



(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

The vectors of nodal distances are given in the following table, together with the absolute differences between each pair of vectors.

The nodal distance between phylogenetic trees (a) and (b) is 9, between (a) and (c) is 12, and between (b) and (c) is 19.

Trees (a) and (b) are thus more similar to each other than they are to tree (c).

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

Algorithm

The nodal distance between two unrooted phylogenetic trees can be obtained by computing the distance between each pair of terminal nodes in each of the trees and then computing the absolute difference between the two vectors of nodal distances.

Algorithm

function nodal distance(T_1, T_2)
    $L \leftarrow$ terminal node labels in T_1 and T_2
    $n \leftarrow length(L)$
    $d \leftarrow 0$
    for $i \leftarrow 1, \ldots, n-1$ **do**
        $i_1 \leftarrow$ terminal node of T_1 labeled $L[i]$
        $i_2 \leftarrow$ terminal node of T_2 labeled $L[i]$
        for $j \leftarrow i+1, \ldots, n$ **do**
            $j_1 \leftarrow$ terminal node of T_1 labeled $L[j]$
            $j_2 \leftarrow$ terminal node of T_2 labeled $L[j]$
            $d_1 \leftarrow distance(T_1, i_1, j_1)$
            $d_2 \leftarrow distance(T_2, i_2, j_2)$
            $d \leftarrow d + |d_1 - d_2|$
    return d

Remark

When the phylogenetic trees are rooted and not fully resolved, 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.