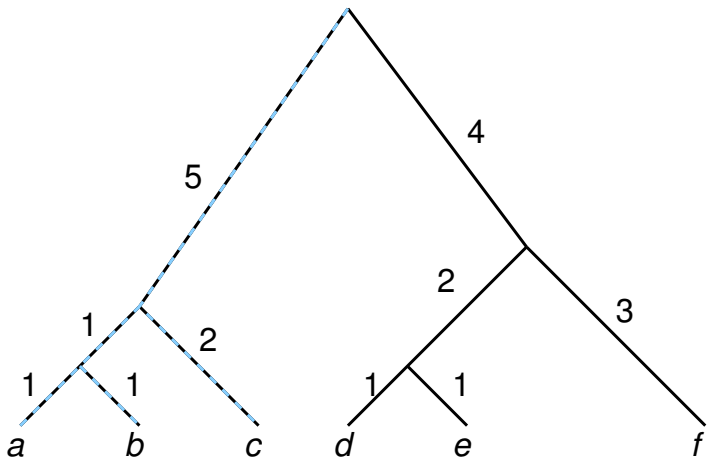
Phylogenetic Trees and k-Dissimilarity Maps

Sven Herrmann
joint work with Katharina T. Huber (UEA), Vincent Moulton (UEA)
and Andreas Spillner (EMAU Greifswald)

School of Computing Sciences
University of East Anglia

Phylogenetics: New data, new Phylogenetic challenges
21 June 2011



Genome Sequences and Phylogenetic Trees

- Given parts of a DNA sequences of different species:

```
GCTTCCA-TCTTGTTATATC
TA-GGCATTGACTAAC-CTG
ACTTATATTGC-TGGGGCCG
```

- Find a suitable phylogenetic tree

- How does one find that tree?

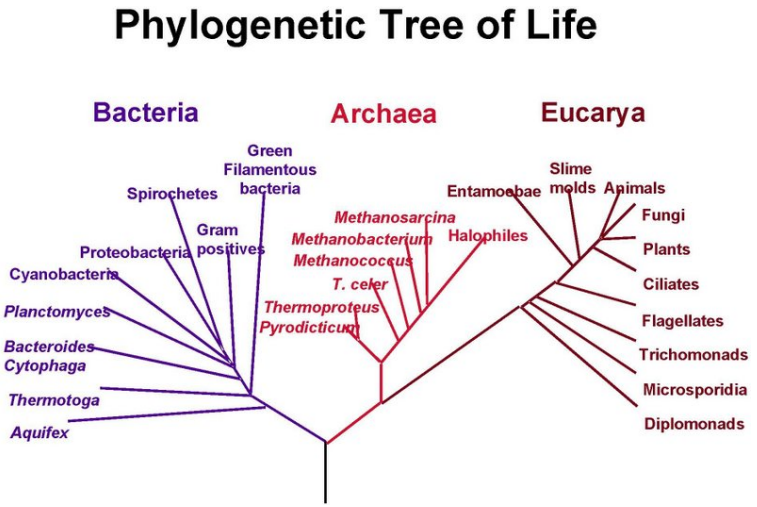
University of East Anglia

Genome Sequences and Phylogenetic Trees

- Given parts of a DNA sequences of different species:

  ```
  GCTTCCA-TCTTGTTATATC
  TA-GGCATTGACTAAC-CTG
  ACTTATATTGC-TGGGGCCG
  ```

- Find a suitable phylogenetic tree



- How does one find that tree?

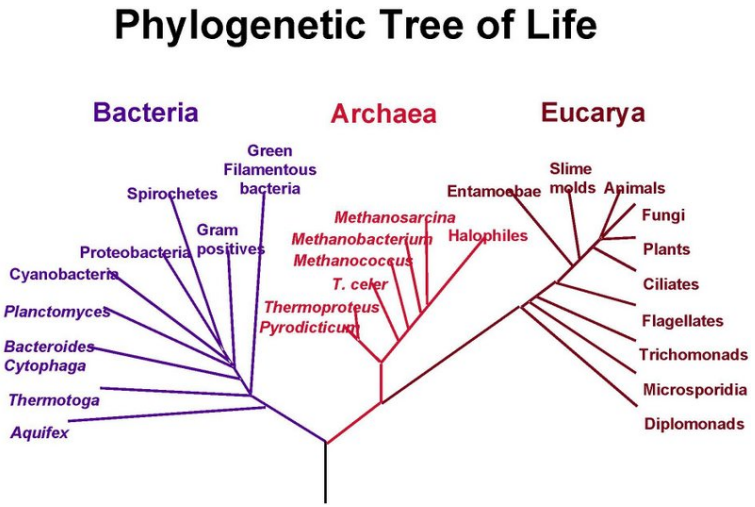
University of East Anglia

Genome Sequences and Phylogenetic Trees

- Given parts of a DNA sequences of different species:

  ```
  GCTTCCA-TCTTGTTATATC
  TA-GGCATTGACTAAC-CTG
  ACTTATATTGC-TGGGGCCG
  ```

- Find a suitable phylogenetic tree



- How does one find that tree?

University of East Anglia

Approach: Use Distances!

- Distance: function $d : X \times X \to \mathbb{R}_{\geq 0}$ assigning a value to each pair of taxa; $d(x, x) = 0$.
- First construct a distance from the data.
- Then construct a tree from the distance.
- Weighted tree $T \implies$ (treelike) distance d_T on the leaves of T.

University of East Anglia

Approach: Use Distances!

- Distance: function $d : X \times X \to \mathbb{R}_{\geq 0}$ assigning a value to each pair of taxa; $d(x, x) = 0$.
- First construct a distance from the data.
- Then construct a tree from the distance.
- Weighted tree $T \implies$ (treelike) distance d_T on the leaves of T.

- e.g. Neighbour-Joining (Saitou & Nei 1987), Neighbour Net (Bryant & Moulton 2004), UPGMA (Sokal & Michener)

Approach: Use Distances!

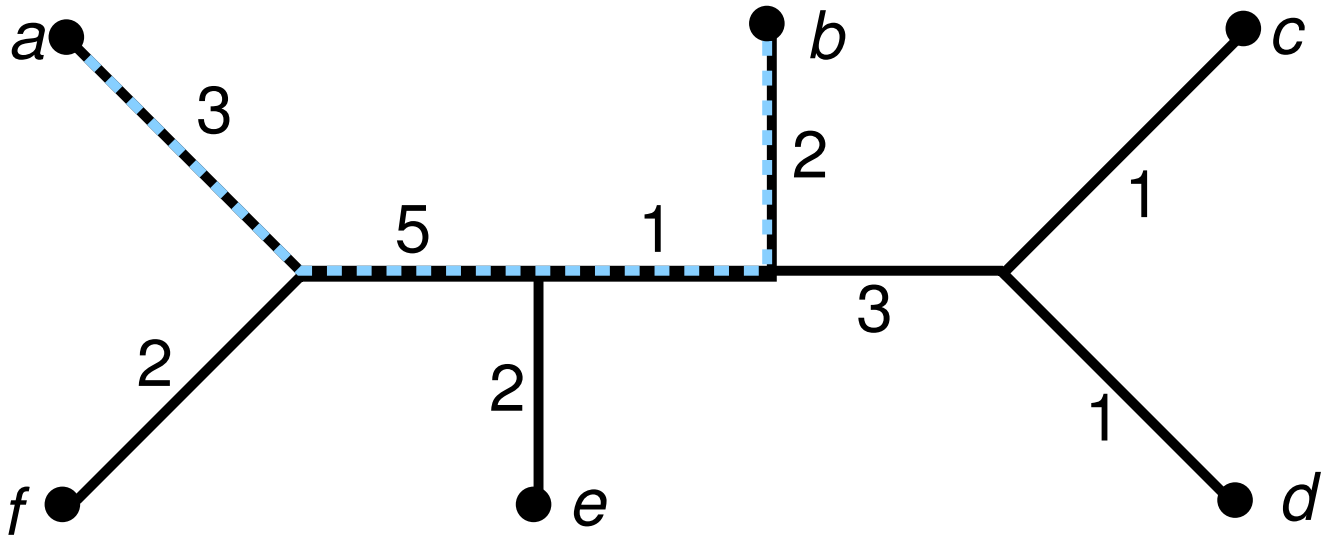
- Distance: function $d : X \times X \to \mathbb{R}_{\geq 0}$ assigning a value to each pair of taxa; $d(x, x) = 0$.
- First construct a distance from the data.
- Then construct a tree from the distance.
- Weighted tree $T \implies$ (treelike) distance d_T on the leaves of T.

- ⇐ Neighbour-Joining (Saitou & Nei 1987)
  Neighbour Net (Bryant & Moulton 2004)
  UPGMA (Sokal & Michener)

University of East Anglia

Approach: Use Distances!

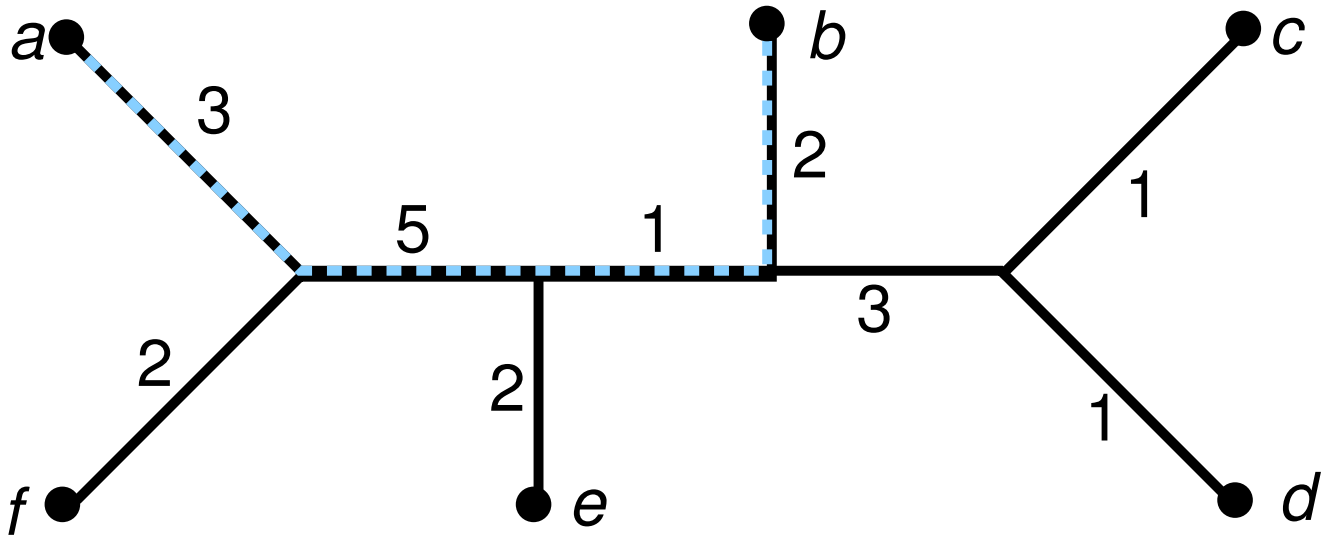
- Distance: function $d : X \times X \to \mathbb{R}_{\geq 0}$ assigning a value to each pair of taxa; $d(x, x) = 0$.
- First construct a distance from the data.
- Then construct a tree from the distance.
- Weighted tree $T \implies$ (treelike) distance d_T on the leaves of T.



- ▸ Neighbour-Joining (Saitou & Nei 1987)
  Neighbour Net (Bryant & Moulton 2004)
  UPGMA (Sokal & Michener)

Approach: Use Distances!

- Distance: function $d : X \times X \to \mathbb{R}_{\geq 0}$ assigning a value to each pair of taxa; $d(x, x) = 0$.
- First construct a distance from the data.
- Then construct a tree from the distance.
- Weighted tree $T \implies$ (treelike) distance d_T on the leaves of T.



-
  - Neighbour-Joining (Saitou & Nei 1987)
    Neighbour Net (Bryant & Moulton 2004)
    UPGMA (Sokal & Michener)

University of East Anglia

4-Point Condition

When is a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ treelike?

Theorem (Buneman 1971)

$d : X \times X \to \mathbb{R}_{\geq 0}$ defines a tree if and only if it satisfies the Triangle Inequality, that is, for all $x, y, y \in X$

$$d(x, y) + d(y, z) \geq d(x, z)\,,$$

and the 4-Point Condition, that is, for all $x, y, z, u \in X$,

$$d(x, y) + d(z, u) \leq \max\left(d(x, z) + d(y, u), d(x, u) + d(y, z)\right)\,.$$

Corollary

There is a polynomial-time algorithm to decide whether a distance d is treelike.

University of East Anglia

4-Point Condition

When is a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ treelike?

Theorem (Buneman 1971)

$d : X \times X \to \mathbb{R}_{\geq 0}$ defines a tree if and only if it satisfies the Triangle Inequality, that is, for all $x, y, y \in X$

$$d(x, y) + d(y, z) \geq d(x, z)\,,$$

and the 4-Point Condition, that is, for all $x, y, z, u \in X$,

$$d(x, y) + d(z, u) \leq \max\left(d(x, z) + d(y, u), d(x, u) + d(y, z)\right)\,.$$

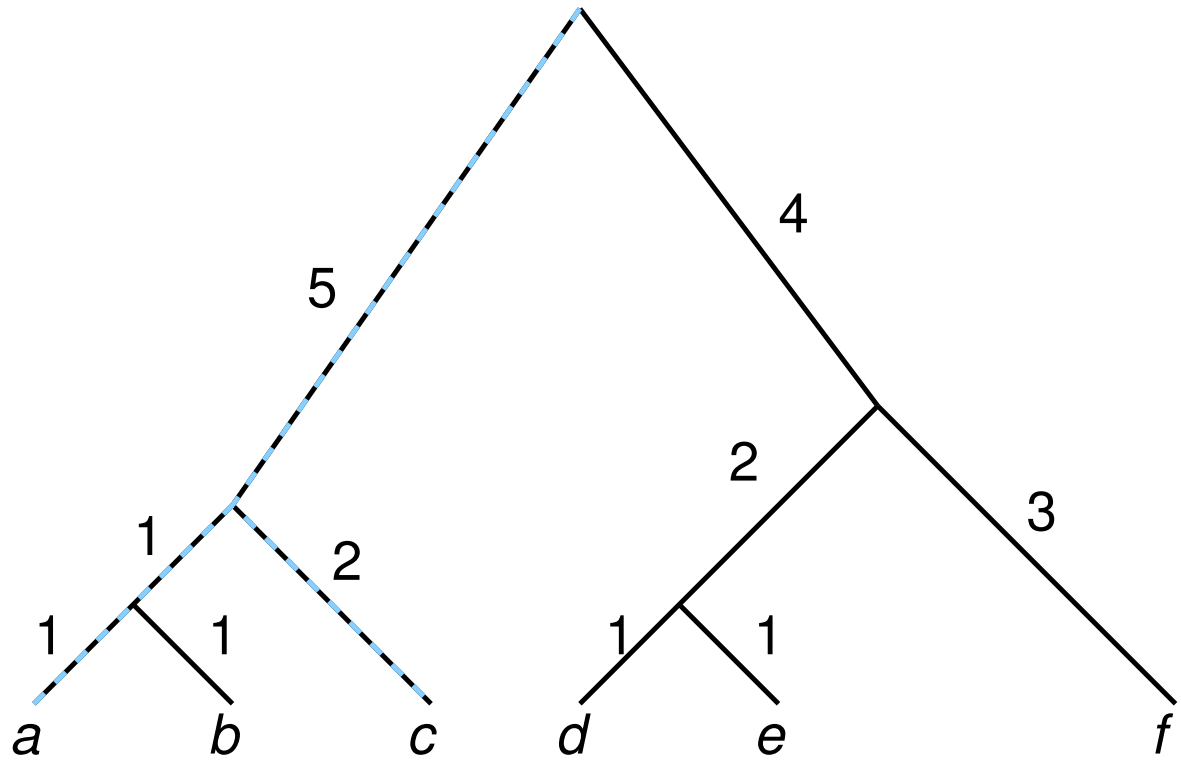
Corollary

There is a polynomial-time algorithm to decide whether a distance d is treelike.

University of East Anglia

Equidistant Trees

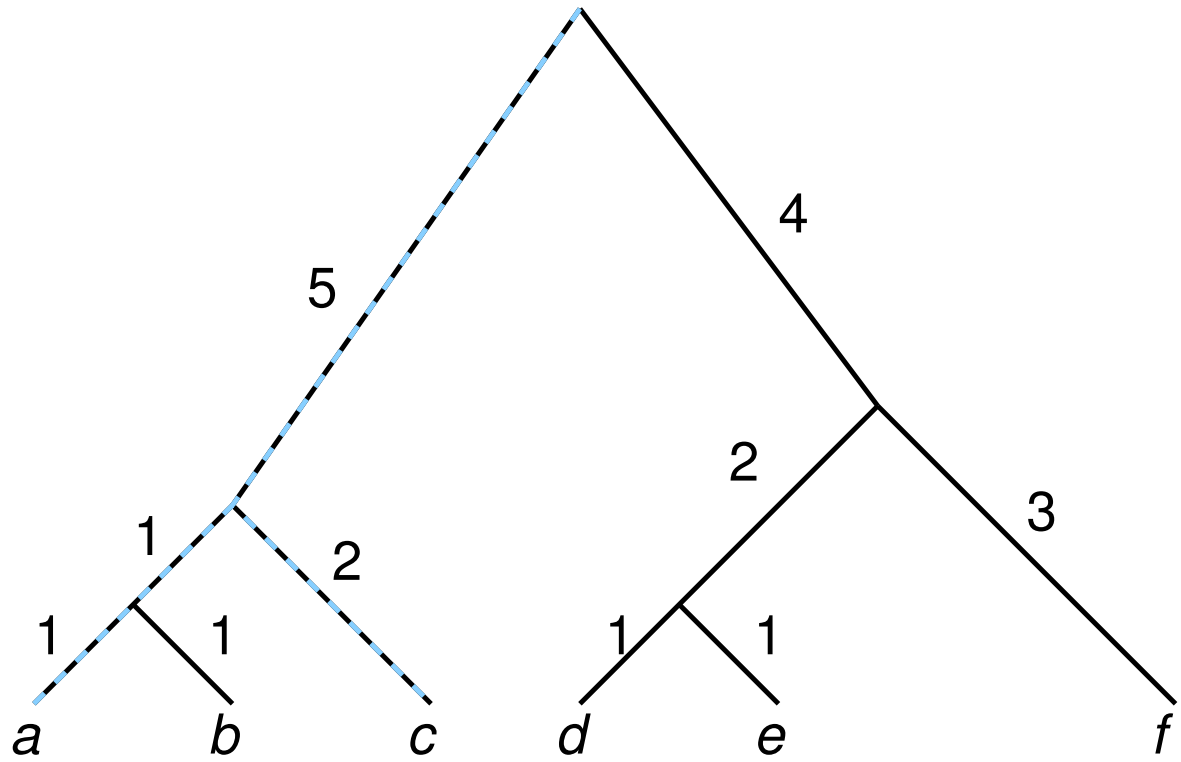
- In a equidistant tree (or heirarchical tree, clocklike tree), the distance from each vertex to a leave does not depend on the path.
- Equidistant weighted tree $T \implies$ (equidistant) distance d_T on the leaves of T.
- When is a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ equidistant?



University of East Anglia

Equidistant Trees

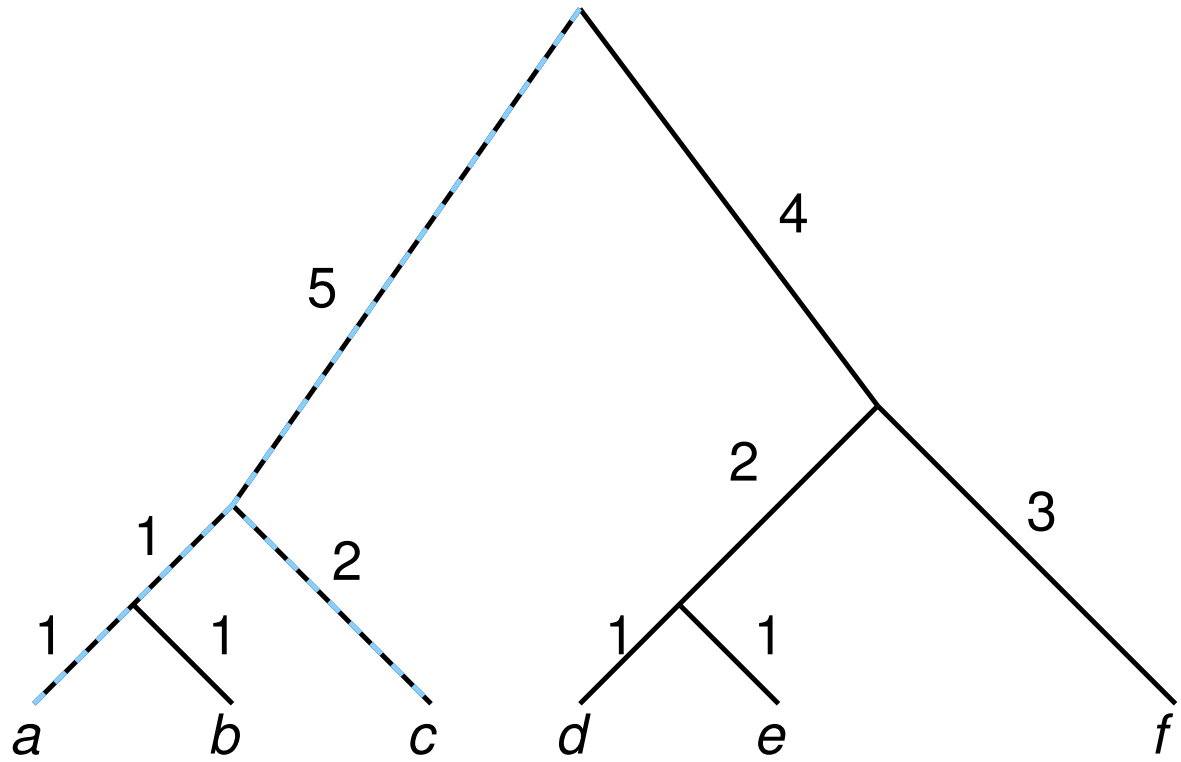
- In a equidistant tree (or heirarchical tree, clocklike tree), the distance from each vertex to a leave does not depend on the path.
- Equidistant weighted tree $T \implies$ (equidistant) distance d_T on the leaves of T.
- When is a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ equidistant?



University of East Anglia

Equidistant Trees

- In a equidistant tree (or heirarchical tree, clocklike tree), the distance from each vertex to a leave does not depend on the path.
- Equidistant weighted tree $T \implies$ (equidistant) distance d_T on the leaves of T.
- When is a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ equidistant?



University of East Anglia

Equidistant Trees

- In a equidistant tree (or heirarchical tree, clocklike tree), the distance from each vertex to a leave does not depend on the path.
- Equidistant weighted tree $T \implies$ (equidistant) distance d_T on the leaves of T.
- When is a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ equidistant?

Theorem

A distance $d : X \times X \to \mathbb{R}_{\geq 0}$ defines an equidistant tree if and only if it satisfies the Ultrametric Condition, that is, for all $x, y, y \in X$

$$d(x, y) \leq \max\left(d(x, z), d(y, z)\right).$$

Corollary

There is a polynomial-time algorithm to decide whether a distance d is equidistant.

University of East Anglia

Equidistant Trees

- In a equidistant tree (or heirarchical tree, clocklike tree), the distance from each vertex to a leave does not depend on the path.
- Equidistant weighted tree $T \implies$ (equidistant) distance d_T on the leaves of T.
- When is a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ equidistant?

Theorem

A distance $d : X \times X \to \mathbb{R}_{\geq 0}$ defines an equidistant tree if and only if it satisfies the Ultrametric Condition, that is, for all $x, y, y \in X$

$$d(x, y) \leq \max\left(d(x, z), d(y, z)\right).$$

Corollary

There is a polynomial-time algorithm to decide whether a distance d is equidistant.

University of East Anglia

Why only use pairs?

- A *k*-dissimilarity map on X is a function $D : \binom{X}{k} \to \mathbb{R}$ assigning a real value to each subset of X with cardinality k.
- Equivalently: totally symmetric function $D : X^k \to \mathbb{R}$.
- There is theoretical work on generalised distances and metrics.
- Several phylogenetic reconstruction approaches use triplets or quartets.
- Pachter and Speyer (2004): *Reconstructing trees from subtree weights*:
- "However, if we are simply given a k-dissimilarity map $D : \binom{X}{k} \to \mathbb{R}$, we do not know how to test whether this map comes from a phylogenetic tree."
- Levy, Yoshida, Pachter (2005): *Beyond pairwise distances: Neighbor-Joining with phylogenetic diversity estimates*:
- "It would be useful to obtain an analog of the four-point condition that characterizes the space of m-dissimilarity maps arising from trees."

University of East Anglia

Why only use pairs?

- A *k*-dissimilarity map on X is a function $D : \binom{X}{k} \to \mathbb{R}$ assigning a real value to each subset of X with cardinality k.
- Equivalently: totally symmetric function $D : X^k \to \mathbb{R}$.
- There is theoretical work on generalised distances and metrics.
- Several phylogenetic reconstruction approaches use triplets or quartets.
- Pachter and Speyer (2004): *Reconstructing trees from subtree weights*:
- "However, if we are simply given a k-dissimilarity map $D : \binom{X}{k} \to \mathbb{R}$, we do not know how to test whether this map comes from a phylogenetic tree."
- Levy, Yoshida, Pachter (2005): *Beyond pairwise distances: Neighbor-Joining with phylogenetic diversity estimates*:
- "It would be useful to obtain an analog of the four-point condition that characterizes the space of m-dissimilarity maps arising from trees."

University of East Anglia

Why only use pairs?

- A *k*-dissimilarity map on X is a function $D : \binom{X}{k} \to \mathbb{R}$ assigning a real value to each subset of X with cardinality k.
- Equivalently: totally symmetric function $D : X^k \to \mathbb{R}$.
- There is theoretical work on generalised distances and metrics.
- Several phylogenetic reconstruction approaches use triplets or quartets.
- Pachter and Speyer (2004): *Reconstructing trees from subtree weights*:
- "However, if we are simply given a k-dissimilarity map $D : \binom{X}{k} \to \mathbb{R}$, we do not know how to test whether this map comes from a phylogenetic tree."
- Levy, Yoshida, Pachter (2005): *Beyond pairwise distances: Neighbor-Joining with phylogenetic diversity estimates*:
- "It would be useful to obtain an analog of the four-point condition that characterizes the space of m-dissimilarity maps arising from trees."

University of East Anglia

Why only use pairs?

- A *k*-dissimilarity map on X is a function $D : \binom{X}{k} \to \mathbb{R}$ assigning a real value to each subset of X with cardinality k.
- Equivalently: totally symmetric function $D : X^k \to \mathbb{R}$.
- There is theoretical work on generalised distances and metrics.
- Several phylogenetic reconstruction approaches use triplets or quartets.
- Pachter and Speyer (2004): *Reconstructing trees from subtree weights*:
- "However, if we are simply given a k-dissimilarity map $D : \binom{X}{k} \to \mathbb{R}$, we do not know how to test whether this map comes from a phylogenetic tree."
- Levy, Yoshida, Pachter (2005): *Beyond pairwise distances: Neighbor-Joining with phylogenetic diversity estimates*:
- "It would be useful to obtain an analog of the four-point condition that characterizes the space of m-dissimilarity maps arising from trees."

University of East Anglia

Why only use pairs?

- A *k*-dissimilarity map on X is a function $D : \binom{X}{k} \to \mathbb{R}$ assigning a real value to each subset of X with cardinality k.
- Equivalently: totally symmetric function $D : X^k \to \mathbb{R}$.
- There is theoretical work on generalised distances and metrics.
- Several phylogenetic reconstruction approaches use triplets or quartets.
- Pachter and Speyer (2004): *Reconstructing trees from subtree weights*:
- "However, if we are simply given a k-dissimilarity map $D : \binom{X}{k} \to \mathbb{R}$, we do not know how to test whether this map comes from a phylogenetic tree."
- Levy, Yoshida, Pachter (2005): *Beyond pairwise distances: Neighbor-Joining with phylogenetic diversity estimates*:
- "It would be useful to obtain an analog of the four-point condition that characterizes the space of m-dissimilarity maps arising from trees."

University of East Anglia

Why only use pairs?

- A k-dissimilarity map on X is a function $D : \binom{X}{k} \to \mathbb{R}$ assigning a real value to each subset of X with cardinality k.
- Equivalently: totally symmetric function $D : X^k \to \mathbb{R}$.
- There is theoretical work on generalised distances and metrics.
- Several phylogenetic reconstruction approaches use triplets or quartets.
- Pachter and Speyer (2004): *Reconstructing trees from subtree weights*:
- "However, if we are simply given a k-dissimilarity map $D : \binom{X}{k} \to \mathbb{R}$, we do not know how to test whether this map comes from a phylogenetic tree."
- Levy, Yoshida, Pachter (2005): *Beyond pairwise distances: Neighbor-Joining with phylogenetic diversity estimates*:
- "It would be useful to obtain an analog of the four-point condition that characterizes the space of m-dissimilarity maps arising from trees."

University of East Anglia

Why only use pairs?

- A *k*-dissimilarity map on X is a function $D : \binom{X}{k} \to \mathbb{R}$ assigning a real value to each subset of X with cardinality k.
- Equivalently: totally symmetric function $D : X^k \to \mathbb{R}$.
- There is theoretical work on generalised distances and metrics.
- Several phylogenetic reconstruction approaches use triplets or quartets.
- Pachter and Speyer (2004): *Reconstructing trees from subtree weights*:
- "However, if we are simply given a k-dissimilarity map $D : \binom{X}{k} \to \mathbb{R}$, we do not know how to test whether this map comes from a phylogenetic tree."
- Levy, Yoshida, Pachter (2005): *Beyond pairwise distances: Neighbor-Joining with phylogenetic diversity estimates*:
- "It would be useful to obtain an analog of the four-point condition that characterizes the space of m-dissimilarity maps arising from trees."

University of East Anglia

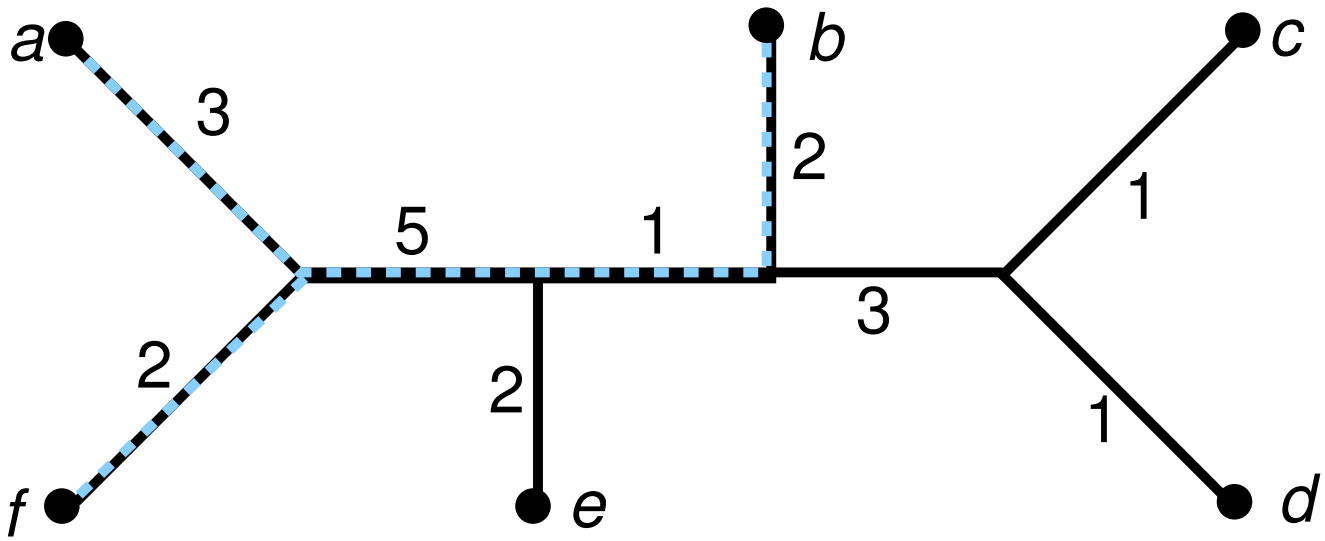
Why only use pairs?

- A *k*-dissimilarity map on X is a function $D : \binom{X}{k} \to \mathbb{R}$ assigning a real value to each subset of X with cardinality k.
- Equivalently: totally symmetric function $D : X^k \to \mathbb{R}$.
- There is theoretical work on generalised distances and metrics.
- Several phylogenetic reconstruction approaches use triplets or quartets.
- Pachter and Speyer (2004): *Reconstructing trees from subtree weights*:
- "However, if we are simply given a k-dissimilarity map $D : \binom{X}{k} \to \mathbb{R}$, we do not know how to test whether this map comes from a phylogenetic tree."
- Levy, Yoshida, Pachter (2005): *Beyond pairwise distances: Neighbor-Joining with phylogenetic diversity estimates*:
- "It would be useful to obtain an analog of the four-point condition that characterizes the space of m-dissimilarity maps arising from trees."

University of East Anglia

Treelike/Equidistant k-Dissimilarity Maps

- Given a tree T one defines D_T^k by assigning to each k-subset $K \subset X$ the total length of the induced subtree.
- $D_T^k(K)$ is also called the phylogenetic diversity of the set K of taxa.

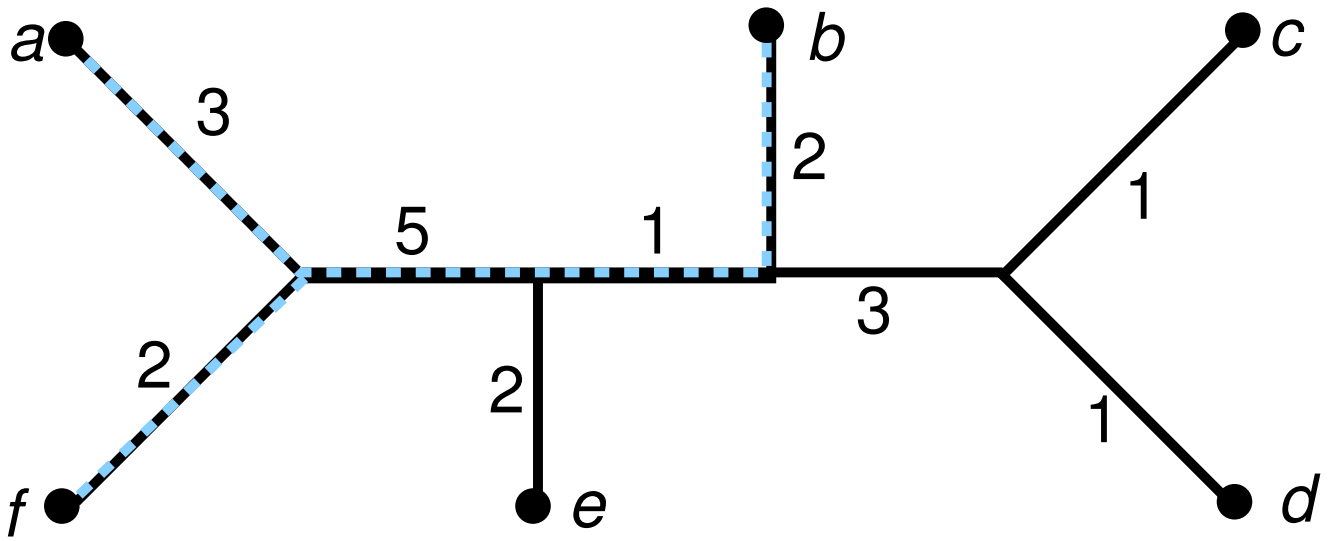


- D is treelike if there exists a phylogenetic tree T such that $D = D_T^k$.

University of East Anglia

Treelike/Equidistant k-Dissimilarity Maps

- Given a tree T one defines D_T^k by assigning to each k-subset $K \subset X$ the total length of the induced subtree.
- $D_T^k(K)$ is also called the phylogenetic diversity of the set K of taxa.

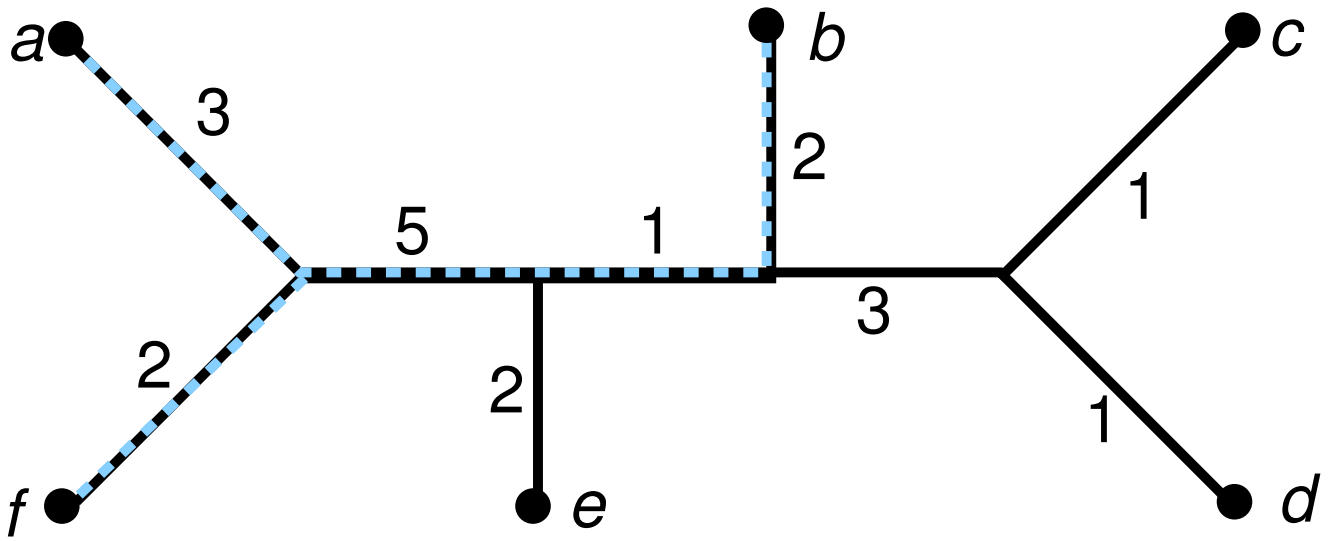


- D is treelike if there exists a phylogenetic tree T such that $D = D_T^k$.

University of East Anglia

Treelike/Equidistant k-Dissimilarity Maps

- Given a tree T one defines D_T^k by assigning to each k-subset $K \subset X$ the total length of the induced subtree.
- $D_T^k(K)$ is also called the phylogenetic diversity of the set K of taxa.

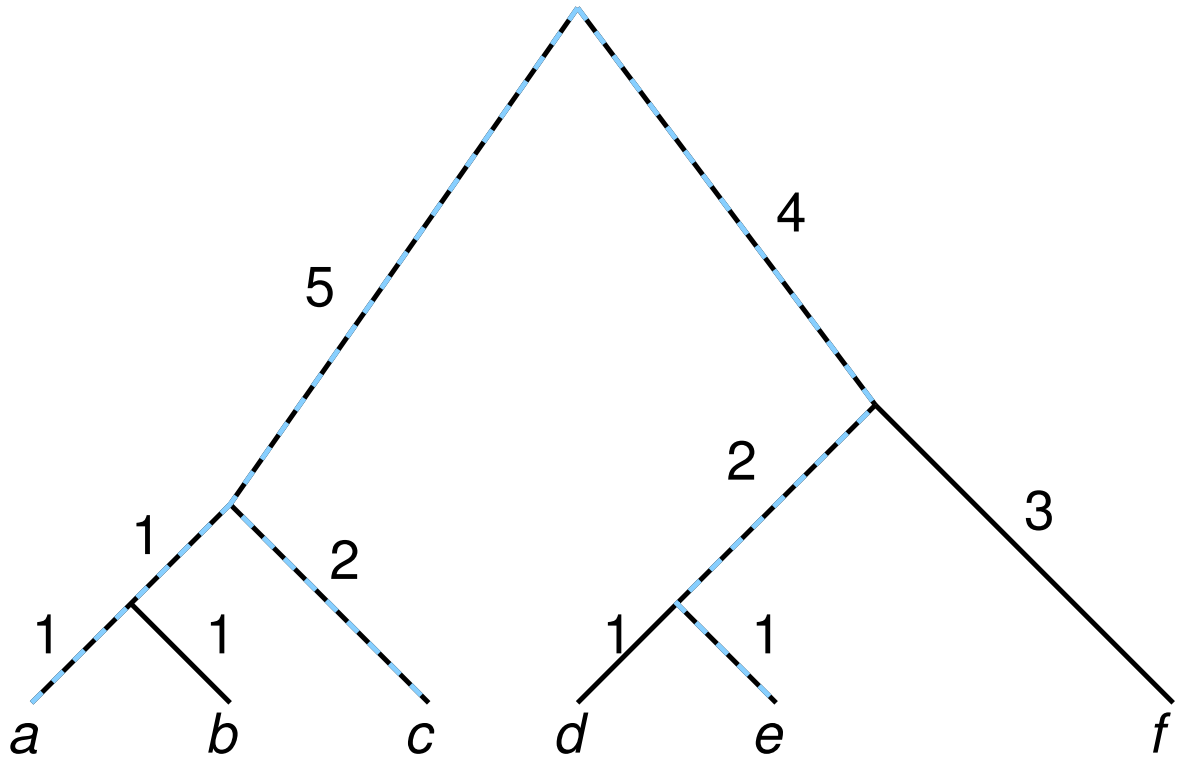


- D is treelike if there exists a phylogenetic tree T such that $D = D_T^k$.

University of East Anglia

Treelike/Equidistant k-Dissimilarity Maps

- Given a tree T one defines D_T^k by assigning to each k-subset $K \subset X$ the total length of the induced subtree.
- $D_T^k(K)$ is also called the phylogenetic diversity of the set K of taxa.



- D is equidistant if there exists a equidistant tree T such that $D = D_T^k$.

University of East Anglia

Existence of $2k$-Point Conditions

Theorem (Huber, Moulton, Spillner, H. 2011)

A k-dissimilarity D on X is treelike/equidistant if and only if the restriction of D to every $2k$-element subset of X is treelike/equidistant.

Corollary

There is a polynomial-time algorithm to decide whether a k-dissimilarity d is treelike/equidistant.

Theorem (Huber, Moulton, Spillner, H. 2011)

For $k \geq 3$ there are non-treelike/equidistant k-dissimilarities whose restriction to every $(2k-1)$-element subset of X is treelike/equidistant.

University of East Anglia

Existence of $2k$-Point Conditions

Theorem (Huber, Moulton, Spillner, H. 2011)

A k-dissimilarity D on X is treelike/equidistant if and only if the restriction of D to every $2k$-element subset of X is treelike/equidistant.

Corollary

There is a polynomial-time algorithm to decide whether a k-dissimilarity d is treelike/equidistant.

Theorem (Huber, Moulton, Spillner, H. 2011)

For $k \geq 3$ there are non-treelike/equidistant k-dissimilarities whose restriction to every $(2k-1)$-element subset of X is treelike/equidistant.

University of East Anglia

Existence of $2k$-Point Conditions

Theorem (Huber, Moulton, Spillner, H. 2011)

A k-dissimilarity D on X is treelike/equidistant if and only if the restriction of D to every $2k$-element subset of X is treelike/equidistant.

Corollary

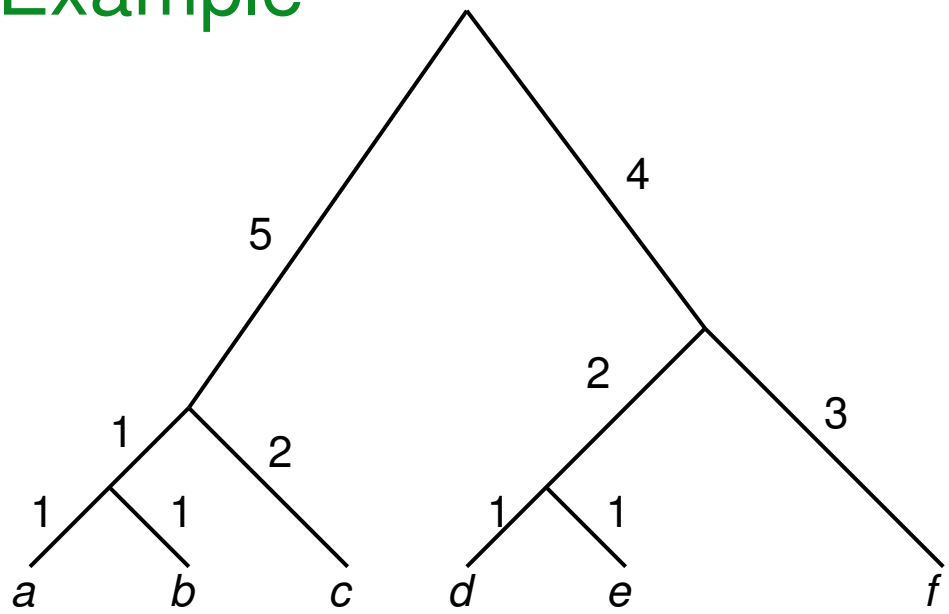
There is a polynomial-time algorithm to decide whether a k-dissimilarity d is treelike/equidistant.

Theorem (Huber, Moulton, Spillner, H. 2011)

For $k \geq 3$ there are non-treelike/equidistant k-dissimilarities whose restriction to every $(2k-1)$-element subset of X is treelike/equidistant.

University of East Anglia

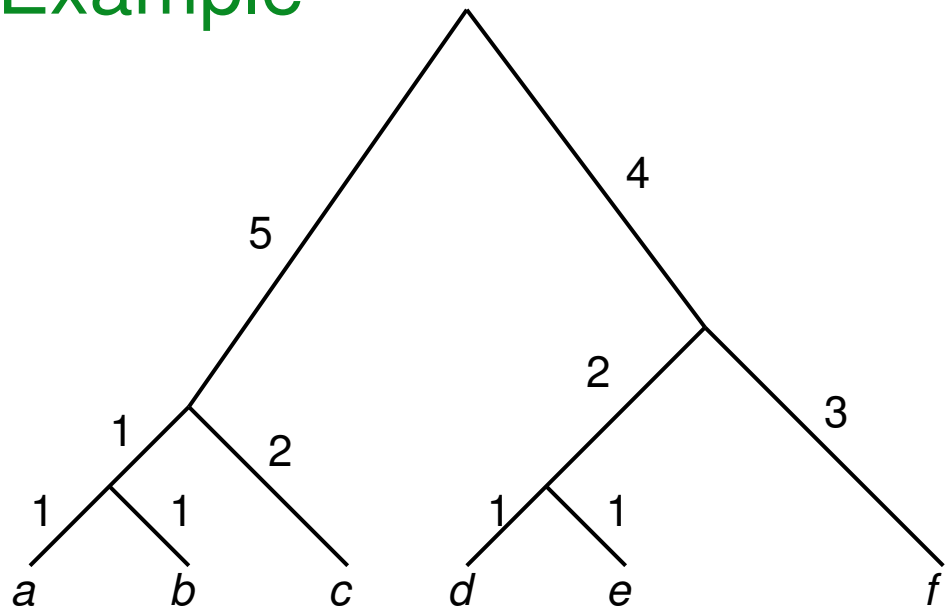
Example

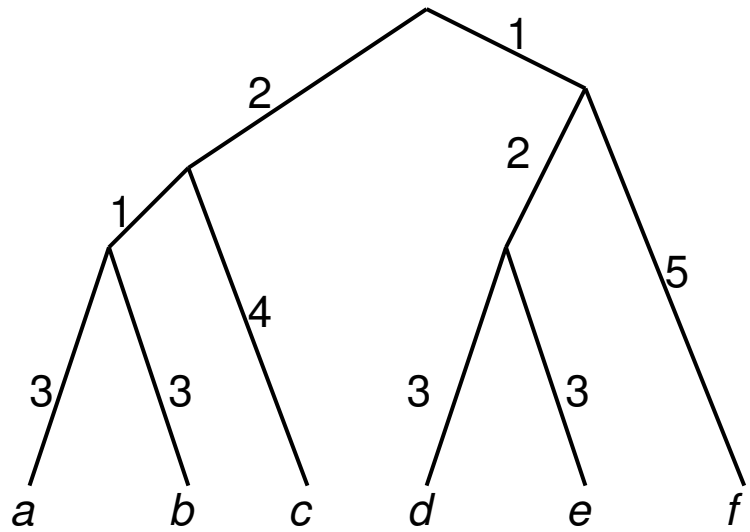


- Let $D(a,b,c) = 13$ and $D(x,y,z) = D^3_T(x,y,z)$ else.
- 5-subset K: if $\{a,b,c\} \subset K$ take right tree else take left tree!
- Restriction of D to every 5-subset is equidistant (treelike).
- It is easily seen, that the left tree is the only one to fit all dissimilarity values despite $D(a,b,c)$.
- However, it can be shown (at least under suitable genericity conditions) that at least the topology can be recovered.

University of East Anglia

Example

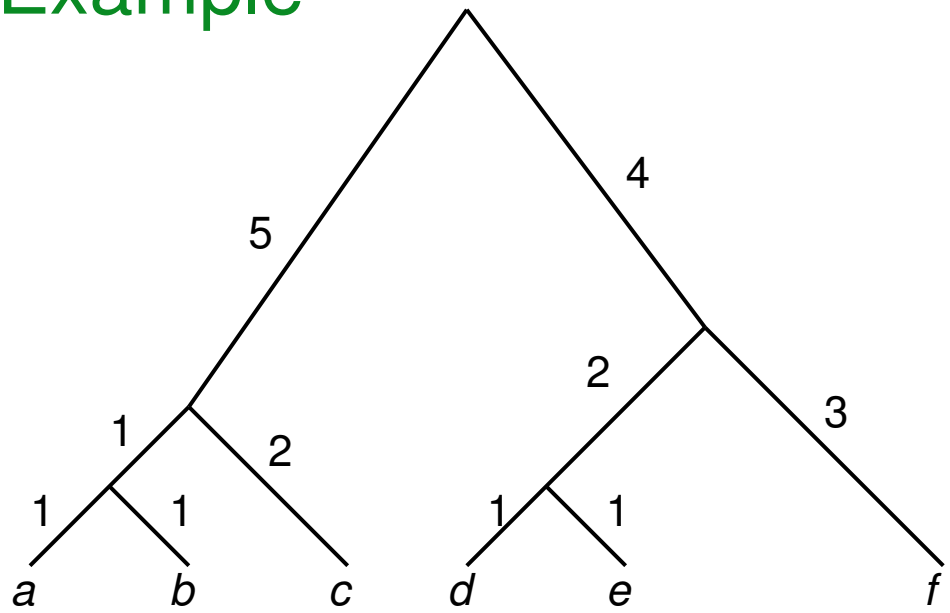


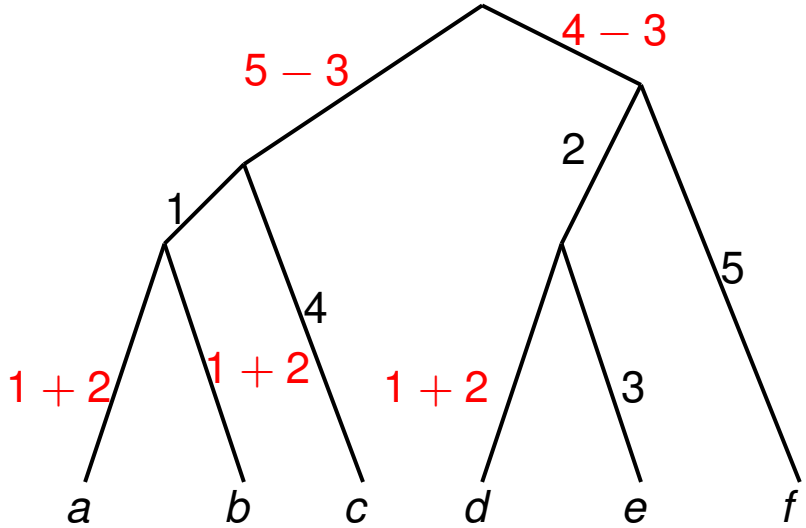


- Let $D(a,b,c) = 13$ and $D(x,y,z) = D^3_T(x,y,z)$ else.
- 5-subset K: if $\{a,b,c\} \subset K$ take right tree else take left tree!
- Restriction of D to every 5-subset is equidistant (treelike).
- It is easily seen, that the left tree is the only one to fit all dissimilarity values despite $D(a,b,c)$.
- However, it can be shown (at least under suitable genericity conditions) that at least the topology can be recovered.

University of East Anglia

Example

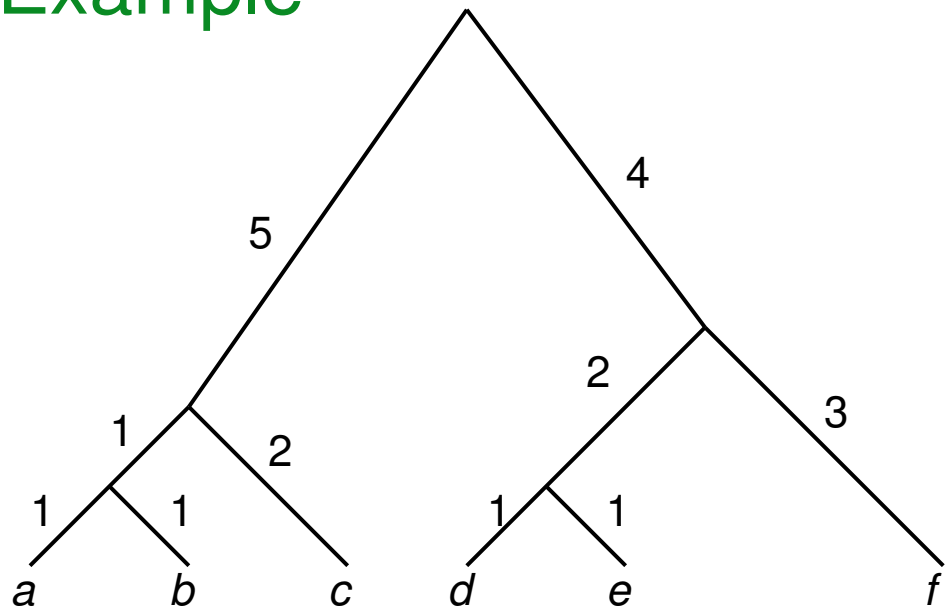


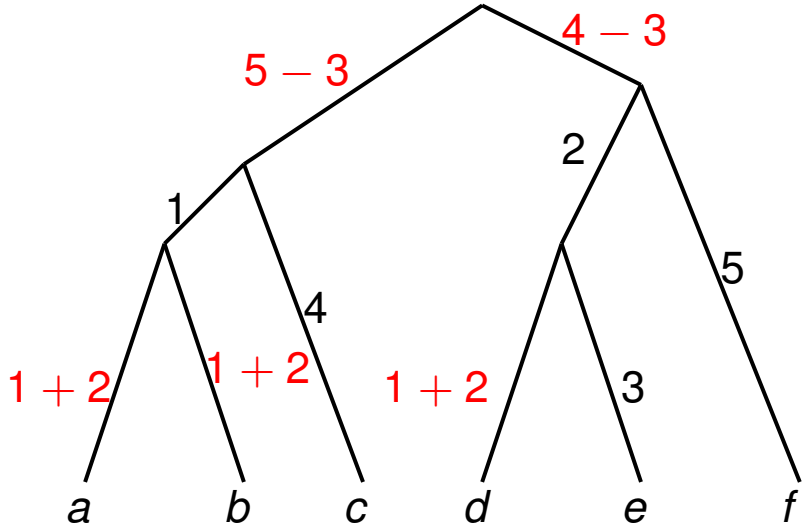


- Let $D(a,b,c) = 13$ and $D(x,y,z) = D^3_T(x,y,z)$ else.
- 5-subset K: if $\{a,b,c\} \subset K$ take right tree else take left tree!
- Restriction of D to every 5-subset is equidistant (treelike).
- It is easily seen, that the left tree is the only one to fit all dissimilarity values despite $D(a,b,c)$.
- However, it can be shown (at least under suitable genericity conditions) that at least the topology can be recovered.

University of East Anglia

Example

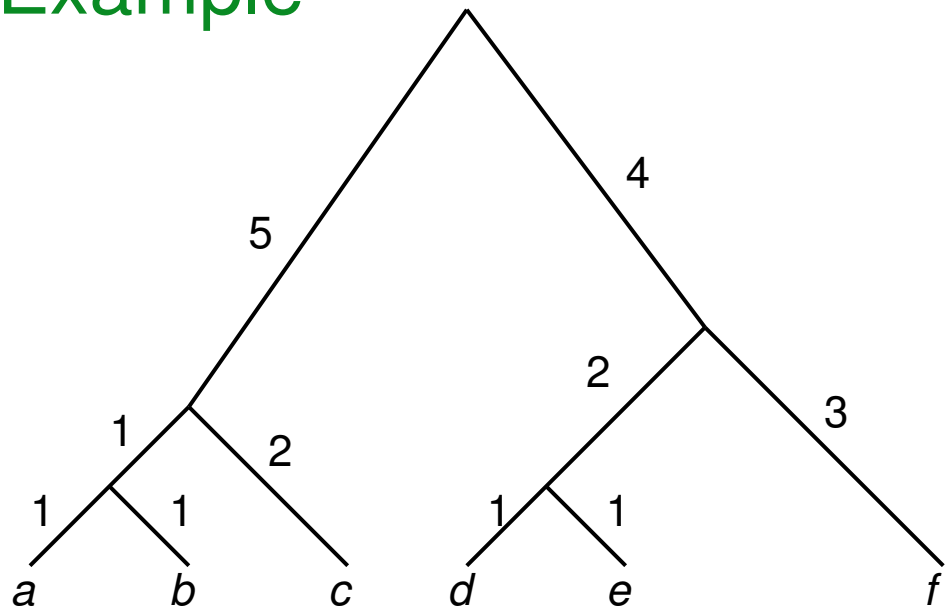


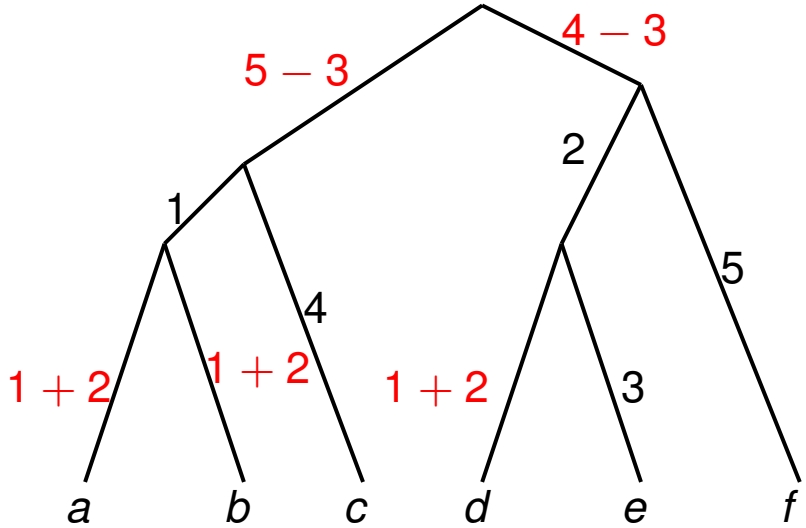


- Let $D(a,b,c) = 13$ and $D(x,y,z) = D^3_T(x,y,z)$ else.
- 5-subset K: if $\{a,b,c\} \subset K$ take right tree else take left tree!
- Restriction of D to every 5-subset is equidistant (treelike).
- It is easily seen, that the left tree is the only one to fit all dissimilarity values despite $D(a,b,c)$.
- However, it can be shown (at least under suitable genericity conditions) that at least the topology can be recovered.

University of East Anglia

Example





- Let $D(a,b,c) = 13$ and $D(x,y,z) = D^3_T(x,y,z)$ else.
- 5-subset K: if $\{a,b,c\} \subset K$ take right tree else take left tree!
- Restriction of D to every 5-subset is equidistant (treelike).
- It is easily seen, that the left tree is the only one to fit all dissimilarity values despite $D(a,b,c)$.
- However, it can be shown (at least under suitable genericity conditions) that at least the topology can be recovered.

University of East Anglia

Three-Dissimilarities

Theorem (Huber, Moulton, Spillner, H. 2011)

A three-dissimilarity D on a set X is equidistant if and only if for all $\{a, b, c, d, e\} \in \binom{X}{5}$

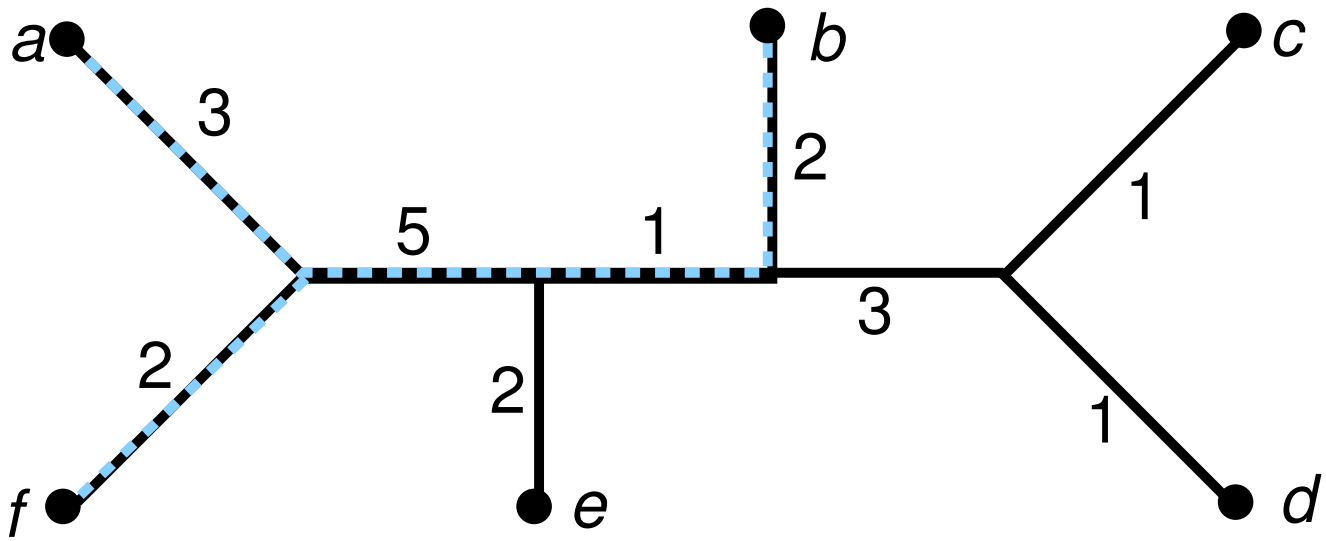
$$D(a,b,e) + D(c,d,e) \leq \max \begin{Bmatrix} D(a,c,e) + D(b,d,e) \\ D(a,d,e) + D(b,c,e) \end{Bmatrix}$$

and for all $\{a, b, c, d, e, e'\} \in \binom{X}{6}$

$$\begin{aligned} & 2D(a,b,e) - D(a,c,e) - D(a,d,e) \\ & - D(b,c,e) - D(b,d,e) + 2D(c,d,e) \\ = \; & 2D(a,b,e') - D(a,c,e') - D(a,d,e') \\ & - D(b,c,e') - D(b,d,e') + 2D(c,d,e') \end{aligned}$$

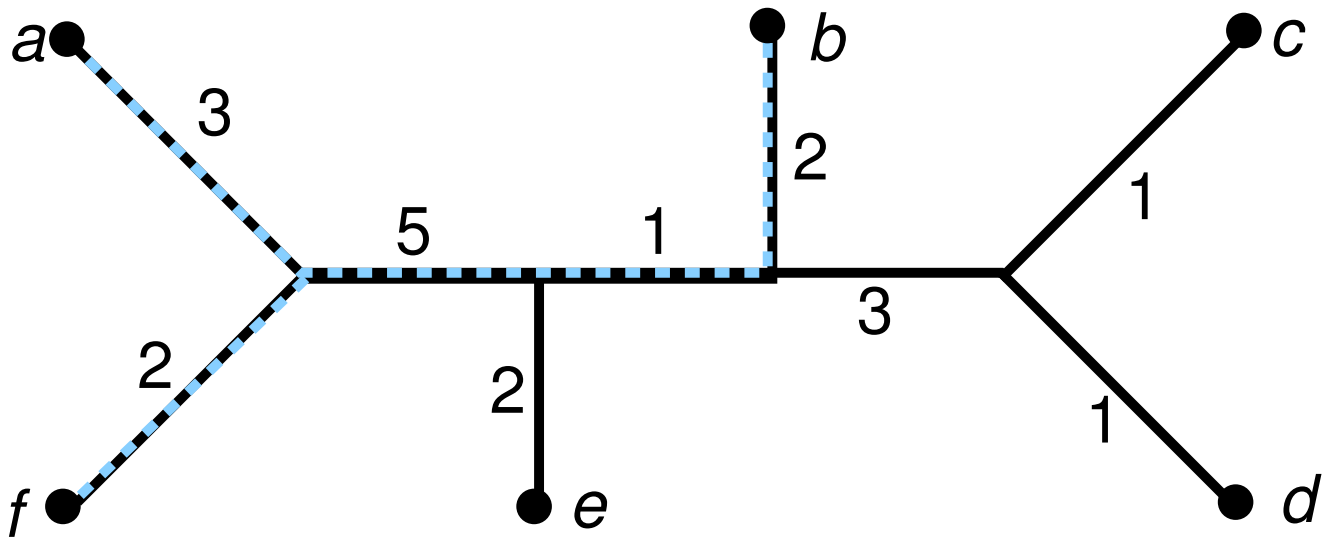
hold.

Thanks for your attention!



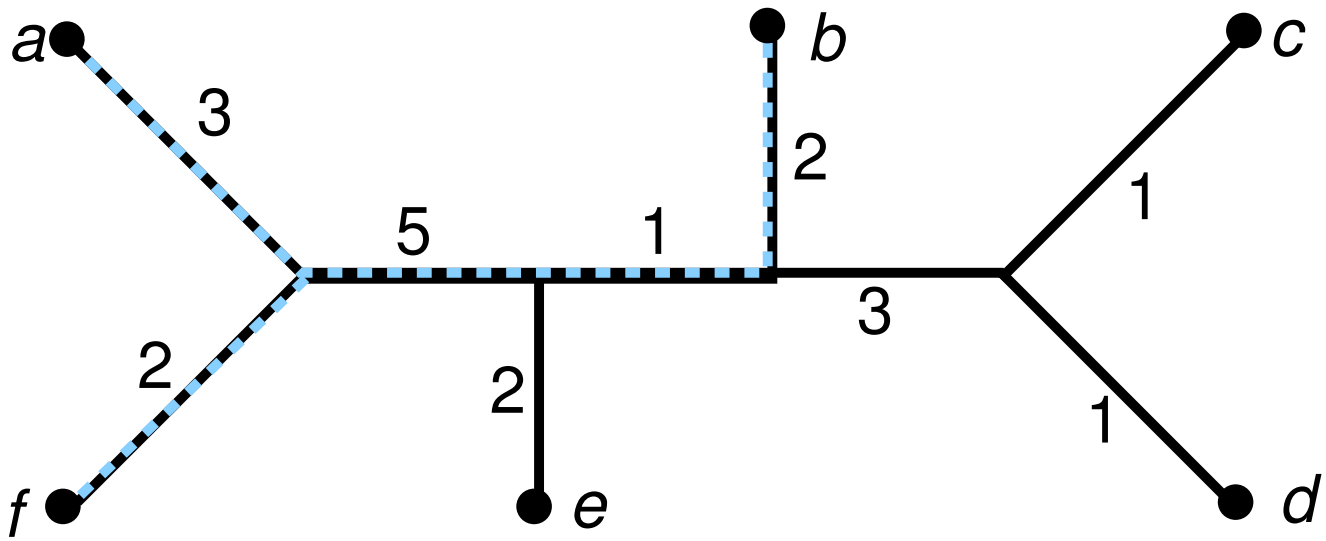
University of East Anglia

Thanks for your attention!



University of East Anglia

Thanks for your attention!



University of East Anglia