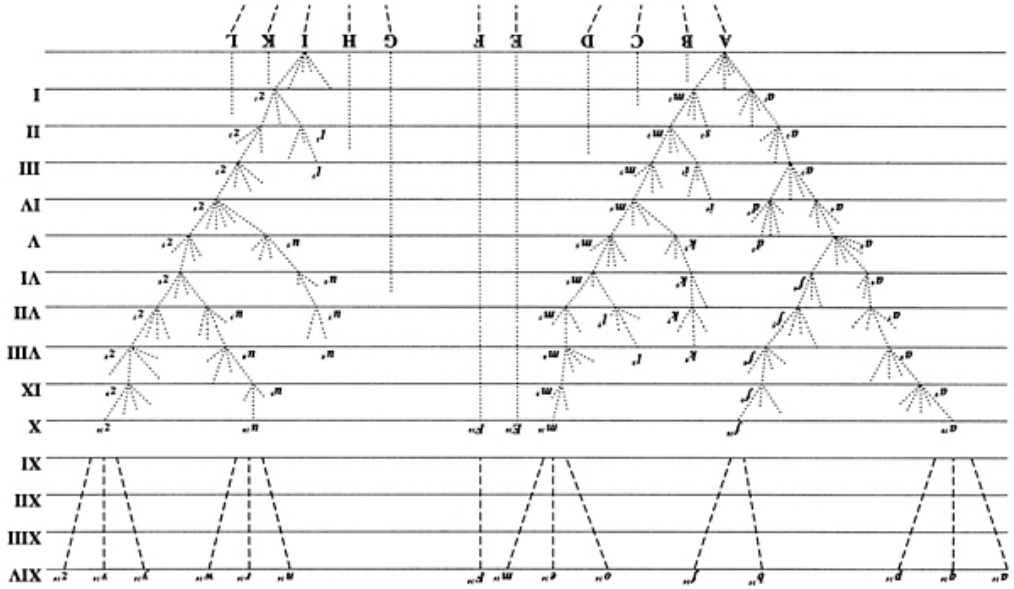
# Identifiability of phylogenetic networks: do not distinguish the indistinguishable

Fabio Pardi[1], Celine Scornavacca[2]

1: CNRS – LIRMM, Montpellier
2: CNRS – ISEM, Montpellier

# Phylogenetic trees

Darwin described evolution as ‘descent with modification’, a phrase that does not necessarily imply a tree representation…

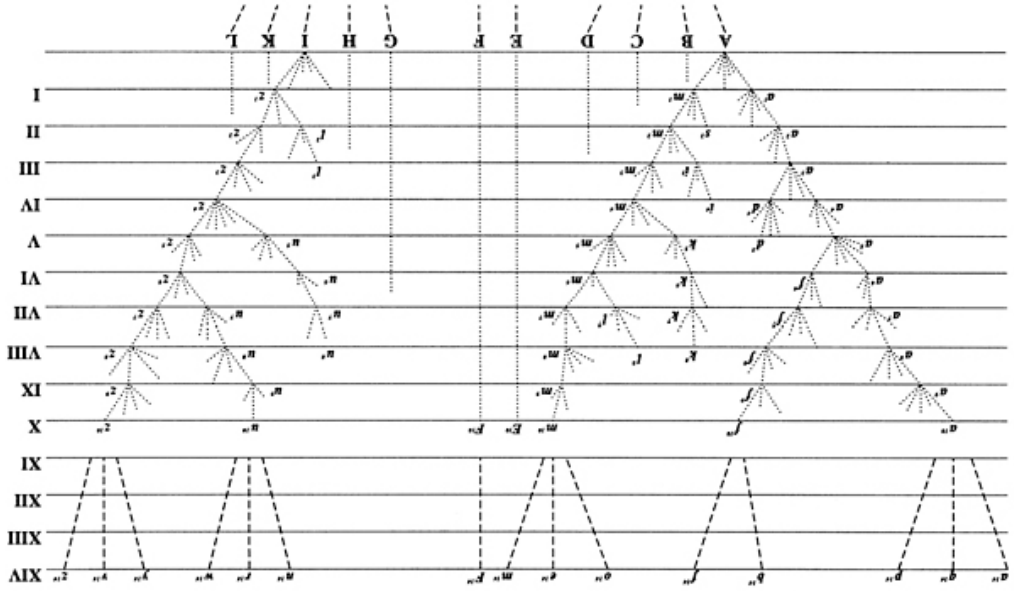



The Origin of Species (1859)

The implicit assumption of using trees is that, at a macroevolutionary scale, each (current or extinct) species or gene *only descends from one ancestor*

# Phylogenetic trees

Darwin described evolution as ‘descent with modification’, a phrase that does not necessarily imply a tree representation…




The Origin of Species (1859)

The implicit assumption of using trees is that, at a macroevolutionary scale, each (current or extinct) species or gene *only descends from one ancestor*

For alleles within a population, we already know this is not true… because of sex (cf. Adam Siepel’s talk yesterday about ARGs)

# Reticulate evolution

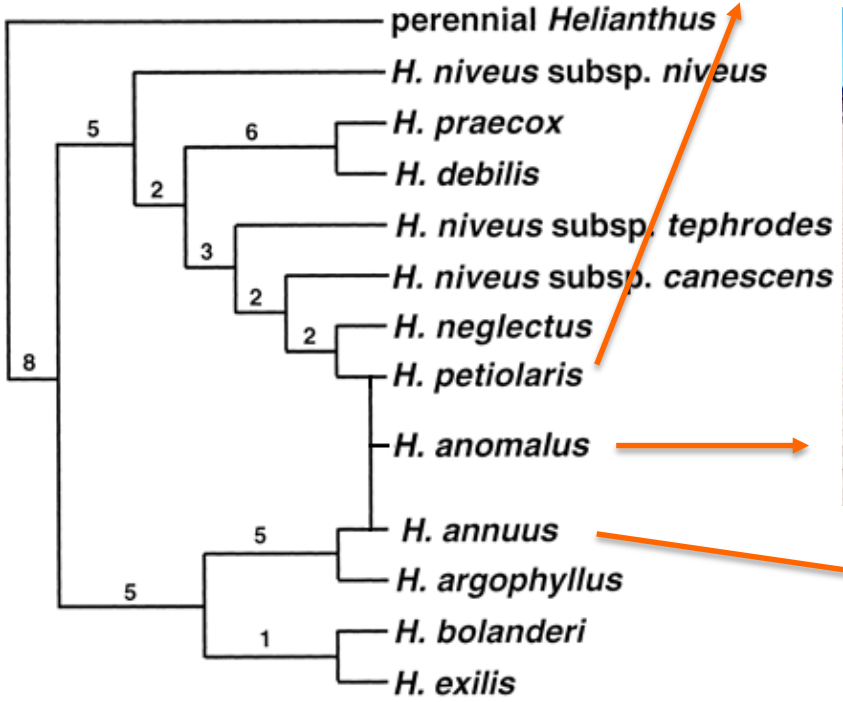
However, sometimes inheritance is from multiple ancestors, because of reticulate events, e.g:

1) Hybrid speciation
2) Lateral gene transfer
3) Recombination

# Reticulate evolution

However, sometimes inheritance is from multiple ancestors, because of reticulate events, e.g:

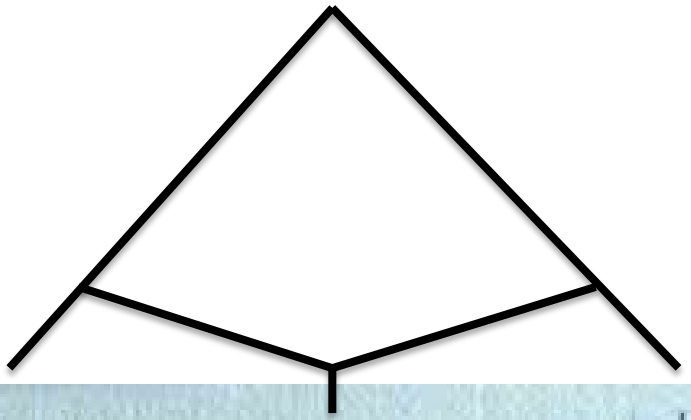
1) **Hybrid speciation**
2) Lateral gene transfer
3) Recombination




# Reticulate evolution

However, sometimes inheritance is from multiple ancestors, because of reticulate events, e.g:

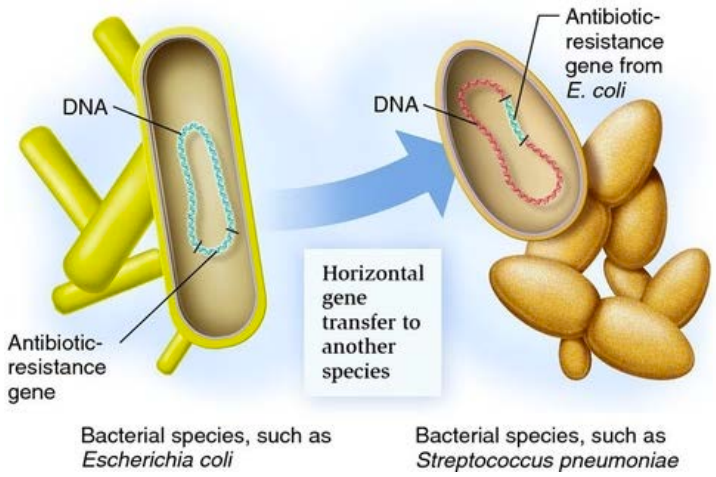
1) **Hybrid speciation**
2) Lateral gene transfer
3) Recombination

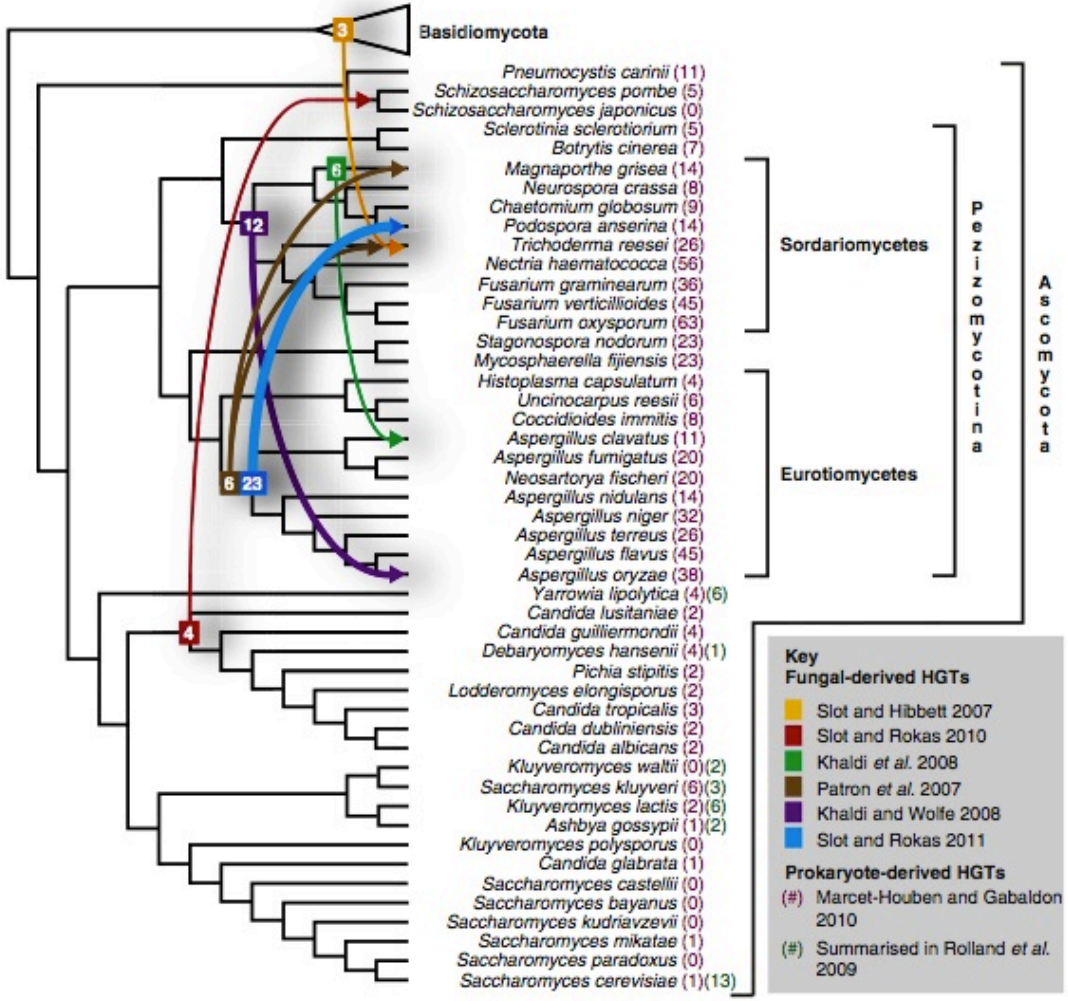


# Reticulate evolution

However, sometimes inheritance is from multiple ancestors, because of reticulate events, e.g:

1) Hybrid speciation
2) **Lateral gene transfer**
3) Recombination

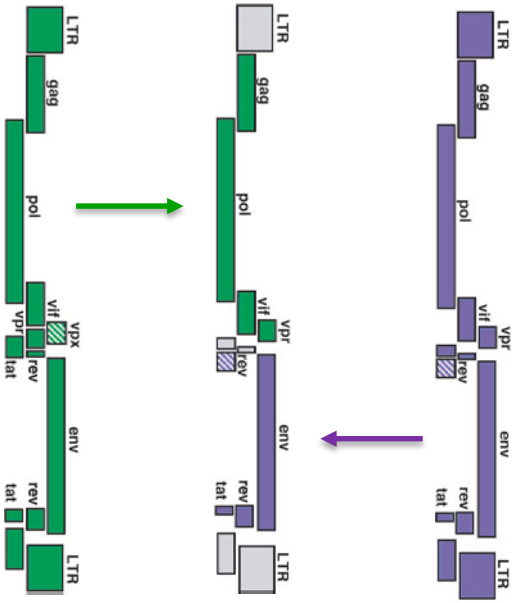


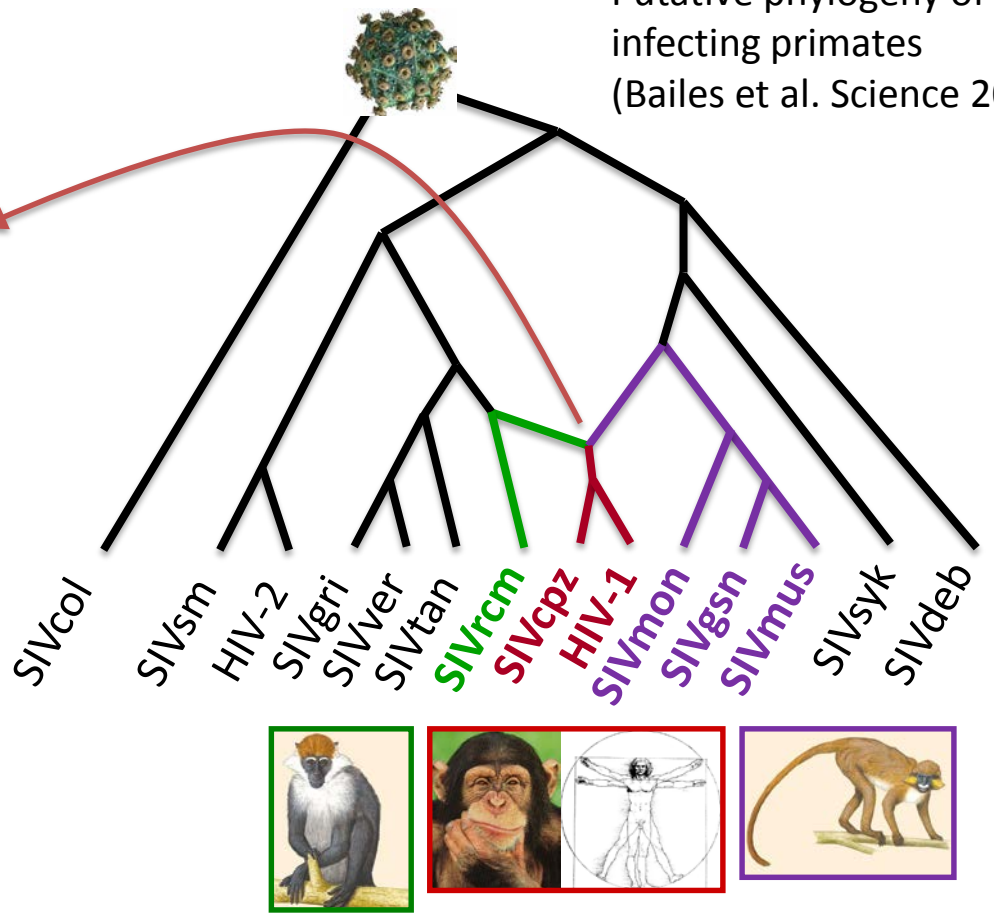
# Reticulate evolution

However, sometimes inheritance is from multiple ancestors, because of reticulate events, e.g:

1) Hybrid speciation
2) Lateral gene transfer
3) **Recombination**



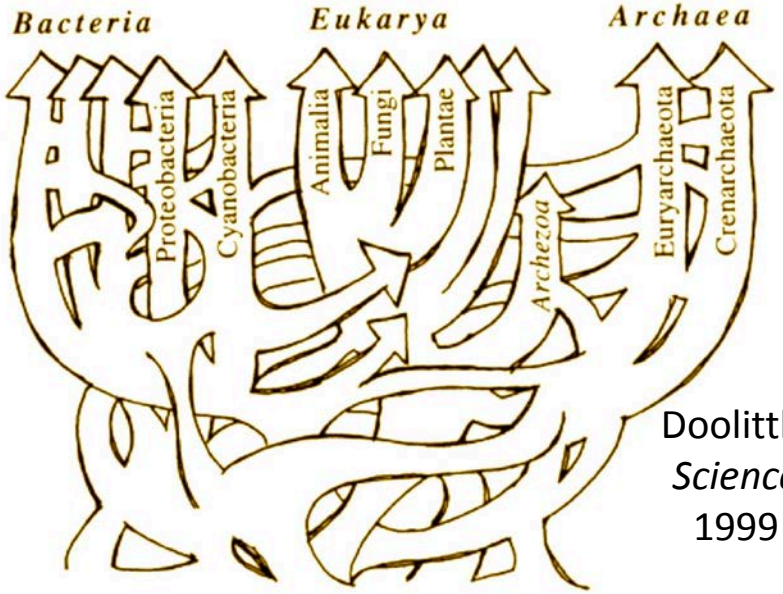

Putative phylogeny of HIV/SIV infecting primates (Bailes et al. Science 2003)



# Phylogenetic networks

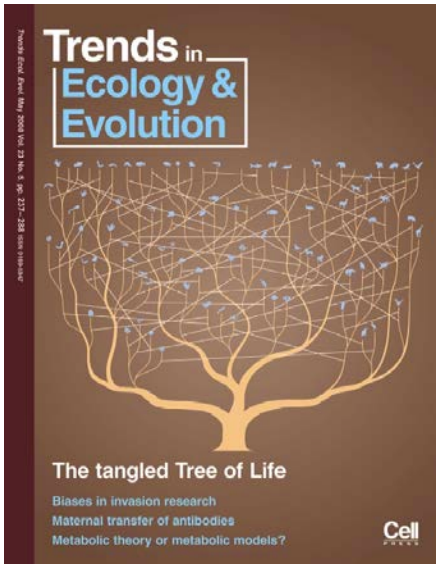
In the presence of reticulate events, phylogenies are networks, not trees

The study of phylogenetic networks is a new interdisciplinary field: maths, CS, biology…

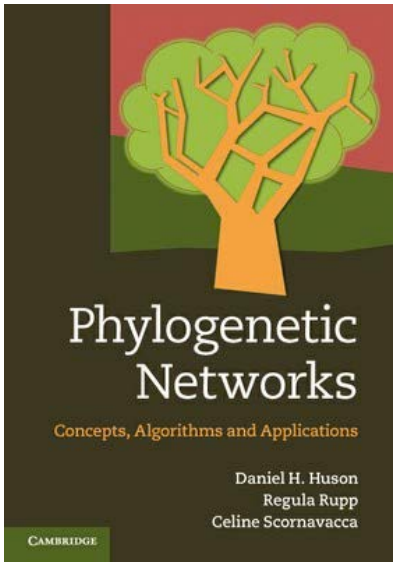



Doolittle
*Science*
1999

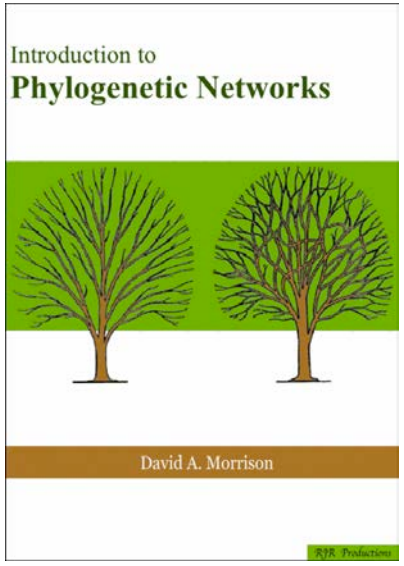
2008



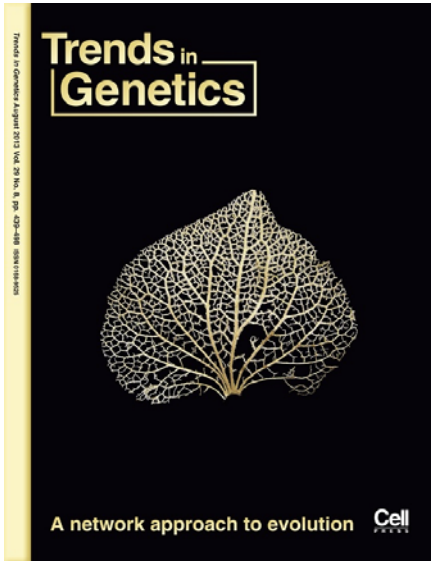

2010




2011

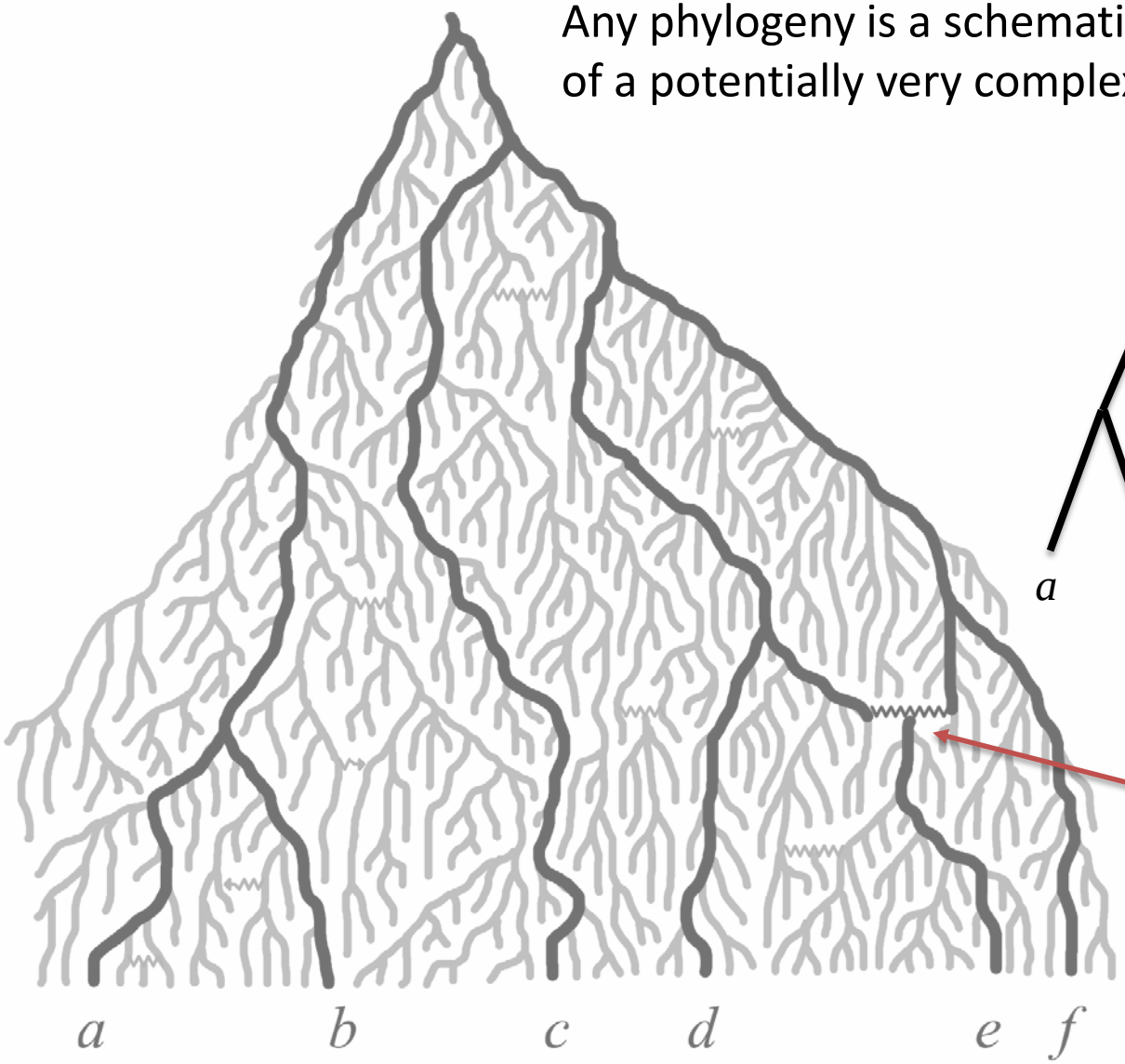


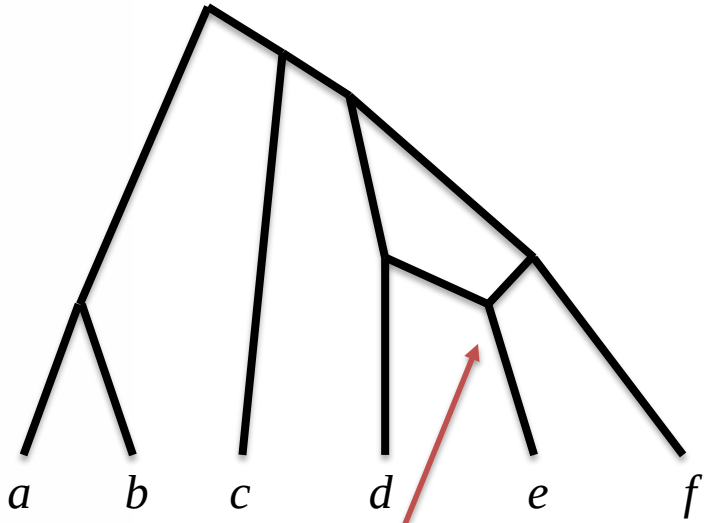

2013



# Phylogenetic networks

Any phylogeny is a schematic representation of part of a potentially very complex story:

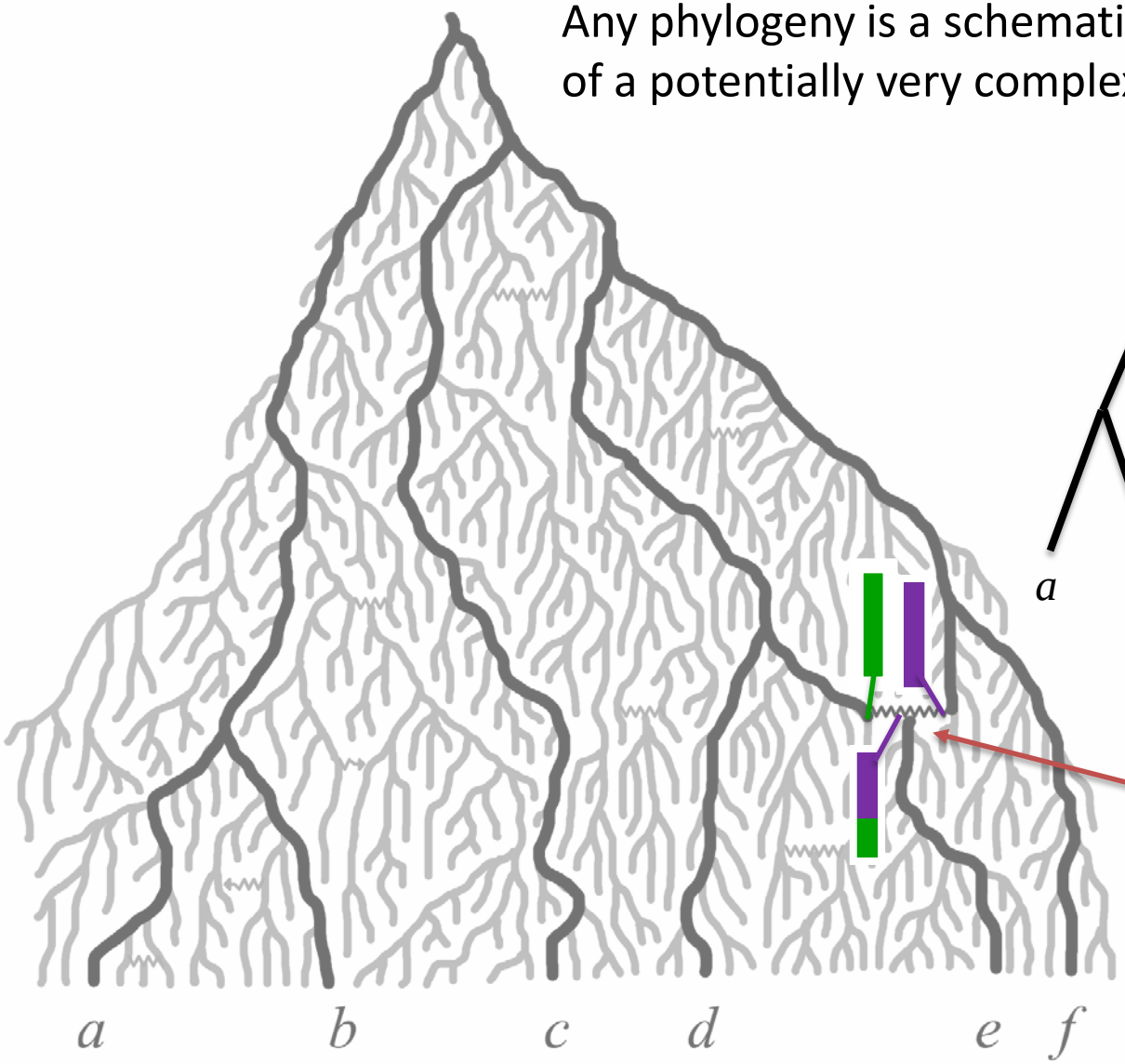


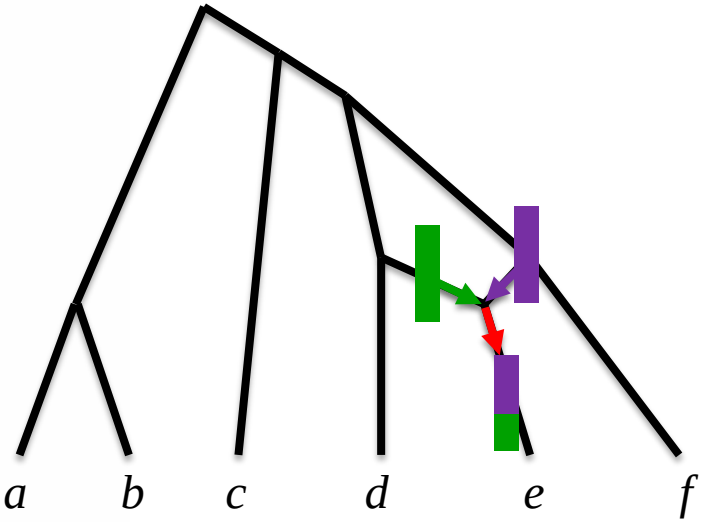




*Reticulation:*
part of the genome is inherited from one lineage, another part from the other

# Phylogenetic networks

Any phylogeny is a schematic representation of part of a potentially very complex story:

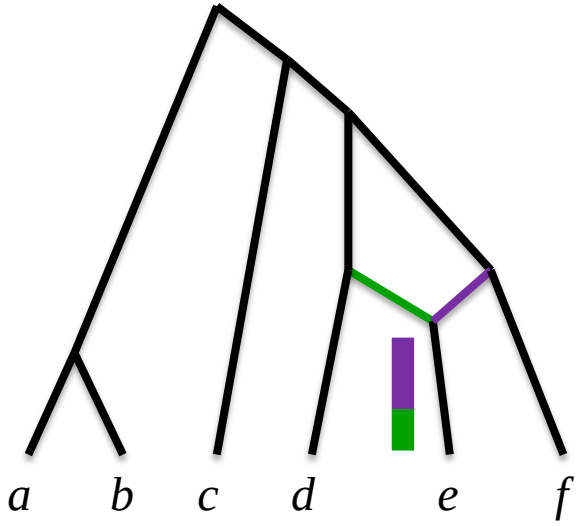



*Reticulation:*
part of the genome is inherited from one lineage, another part from the other

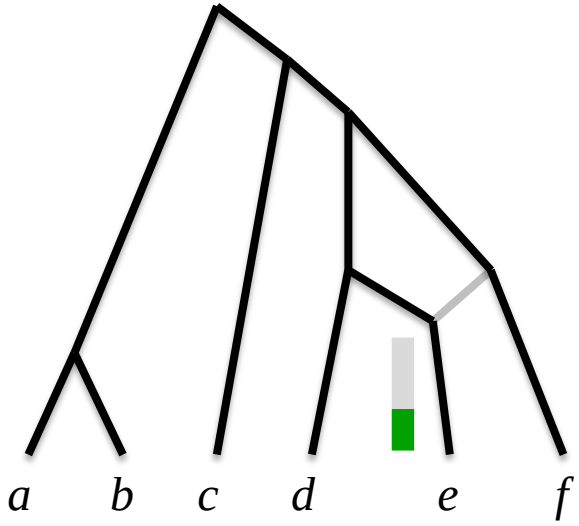
# Trees displayed by a network

Although the evolution of these genomes is best described by a network, the evolution of each part still follows a tree:

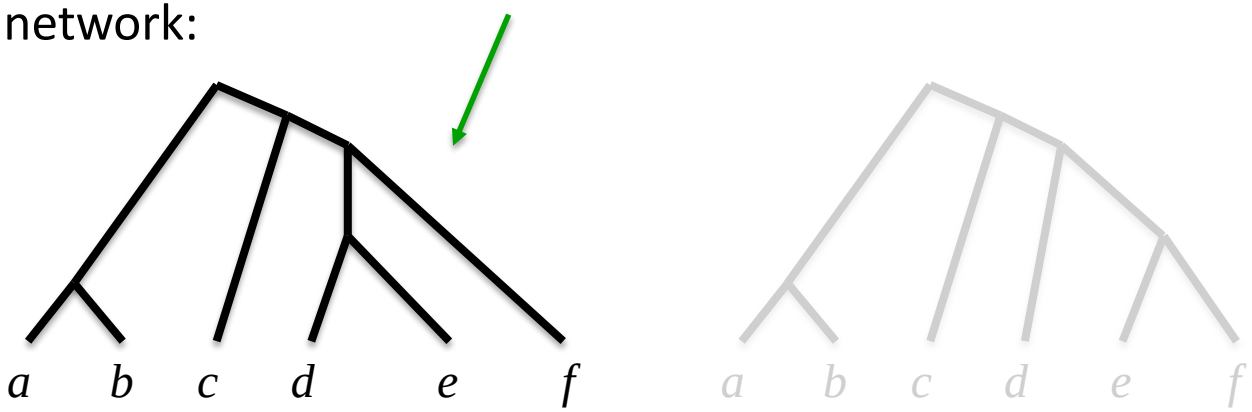


# Trees displayed by a network

Although the evolution of these genomes is best described by a network, the evolution of each part still follows a tree:

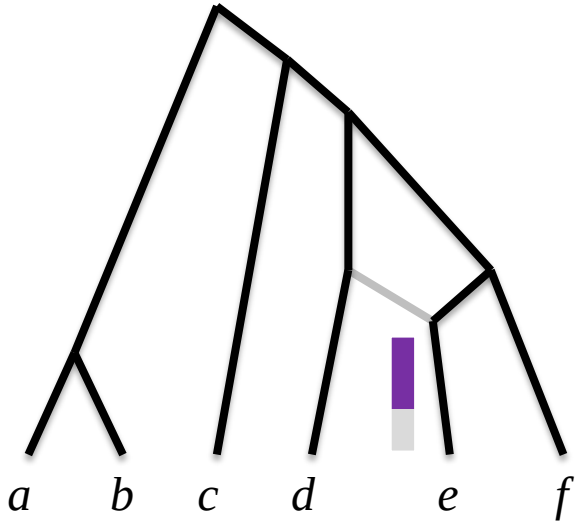



Trees *displayed* by this network:



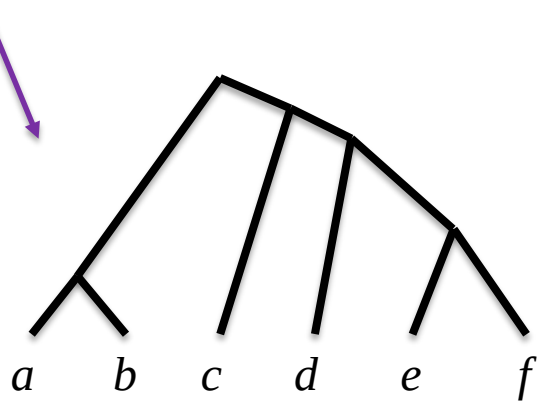
# Trees displayed by a network

Although the evolution of these genomes is best described by a network, the evolution of each part still follows a tree:



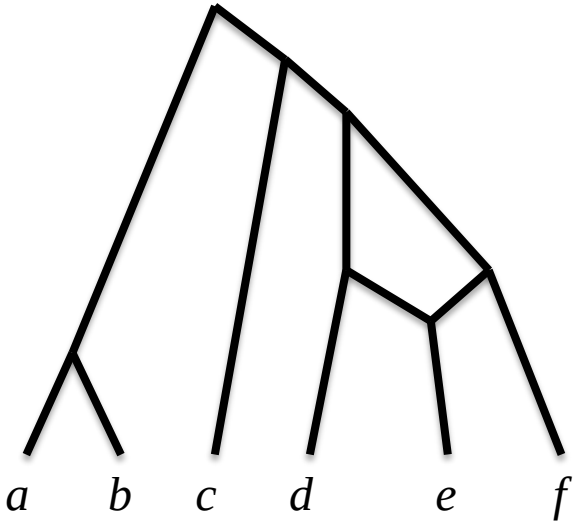

Trees *displayed* by this network:



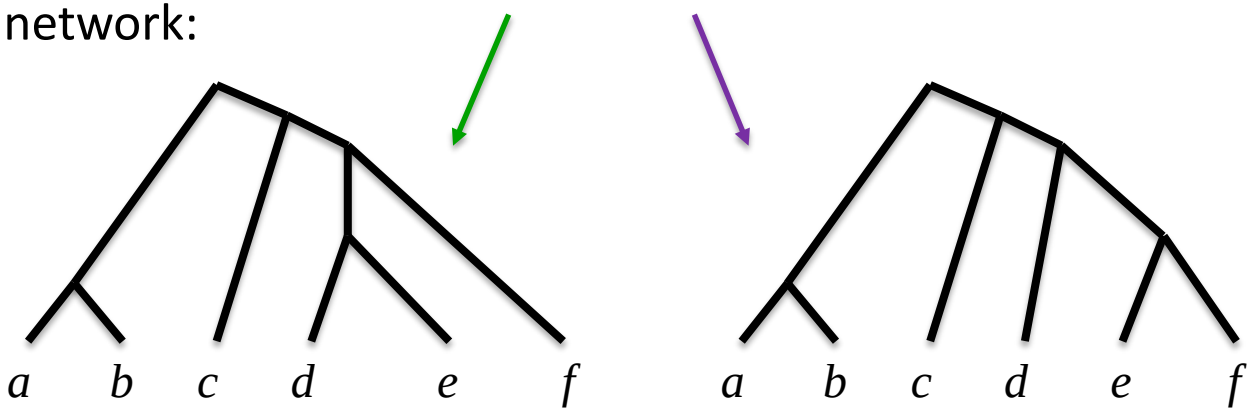



# Trees displayed by a network

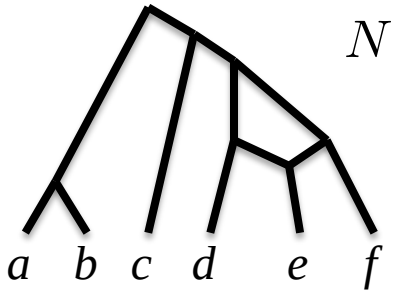
Although the evolution of these genomes is best described by a network, the evolution of each part still follows a tree:




Trees *displayed* by this network:

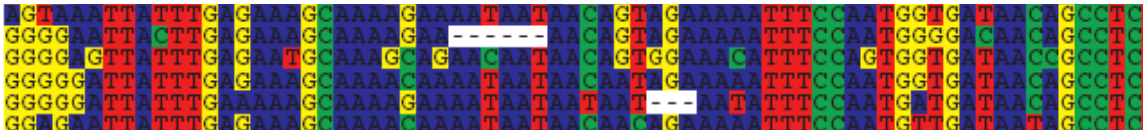


# Phylogenetic network inference



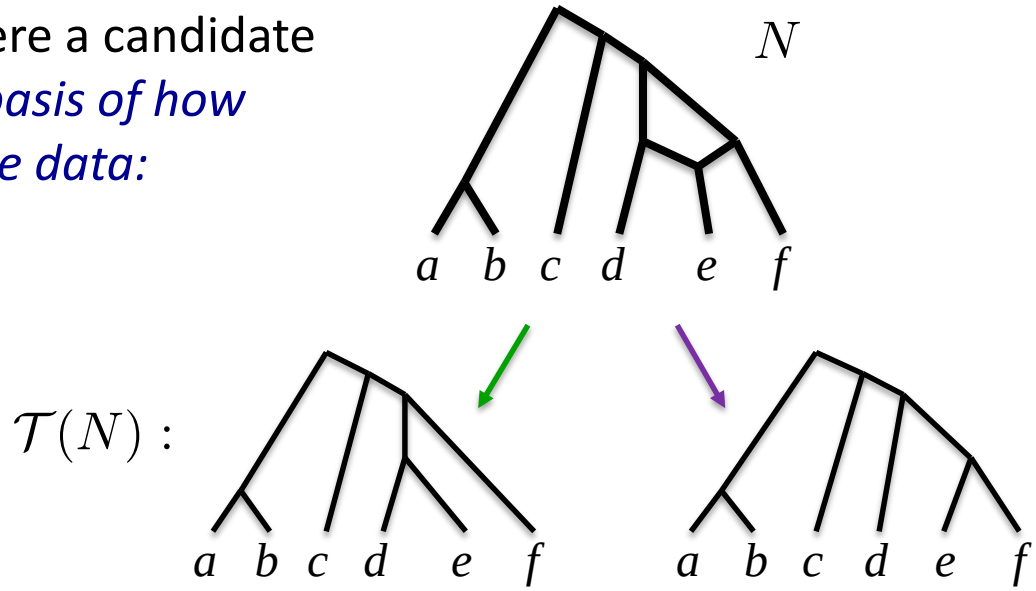

Implicit assumption/hope in the phylogenetic network community: at a macroevolutionary scale, the ratio data/reticulations is ‘large enough’ to allow the inference of the network itself…
*(c.f. ARGs)*

?



# Phylogenetic network inference

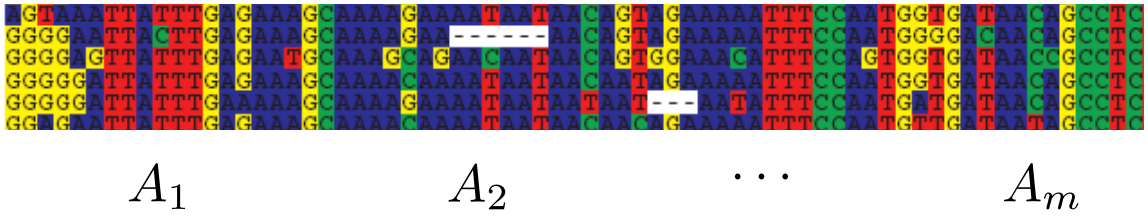
An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data:*




*Many possible formulations:*

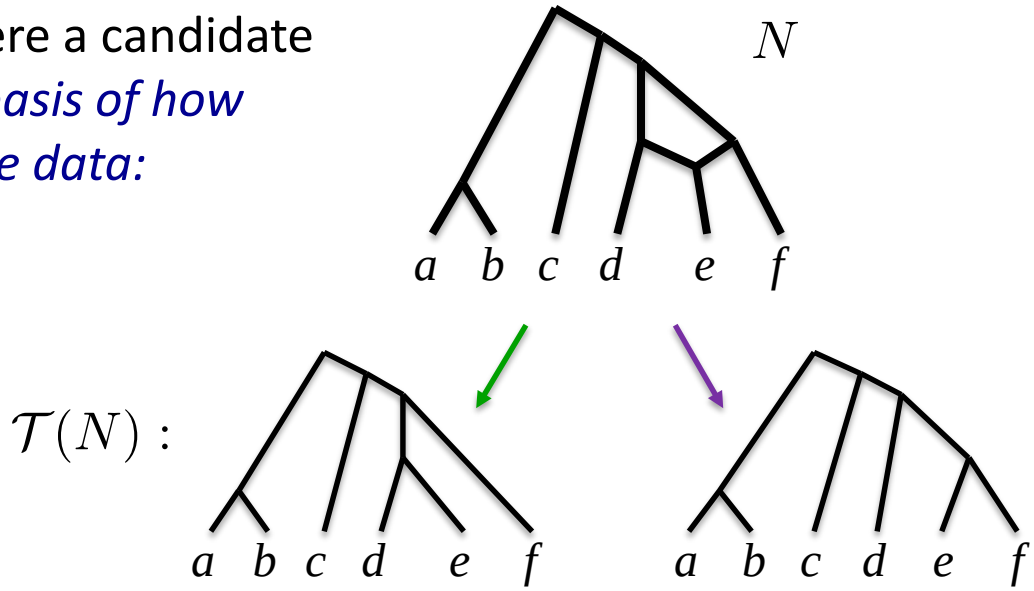
**Data:**

Sequence alignments:
(typically given in blocks)



# Phylogenetic network inference

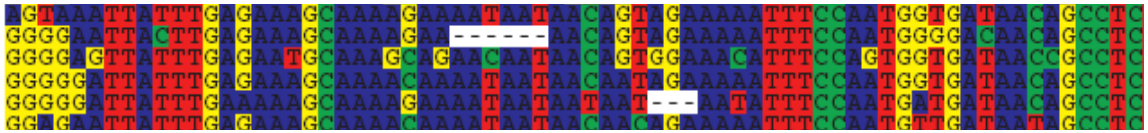
An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data:*




*Many possible formulations:*

**Data:**

Sequence alignments:
(typically given in blocks)



$A_1$ $A_2$ $\cdots$ $A_m$

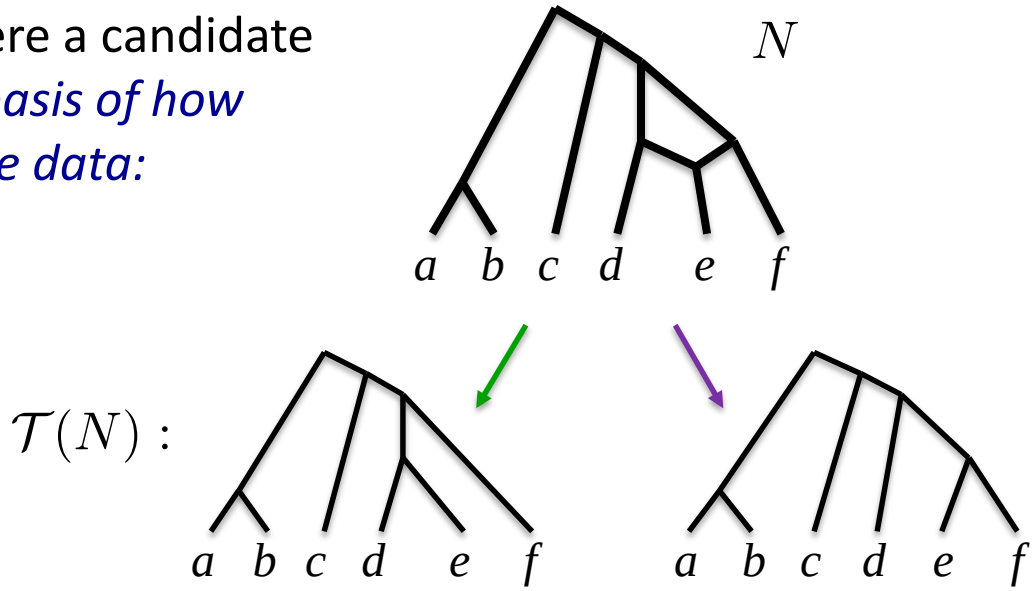
**Goal:**

Find $N$ that minimizes $\boldsymbol{F}(N|A_1, A_2, \ldots, A_m) = \sum_{i=1}^{m} \min_{T \in \mathcal{T}(N)} F(T|A_i)$

subject to constraints on the complexity of $N$

# Phylogenetic network inference

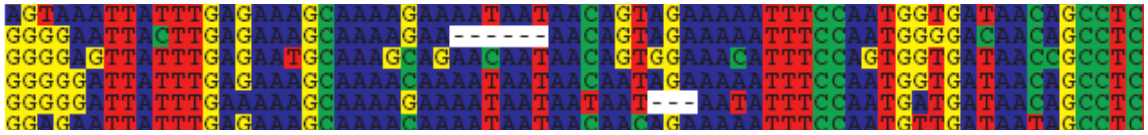
An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data:*




*Many possible formulations:*

**Data:**

Sequence alignments:
(typically given in blocks)



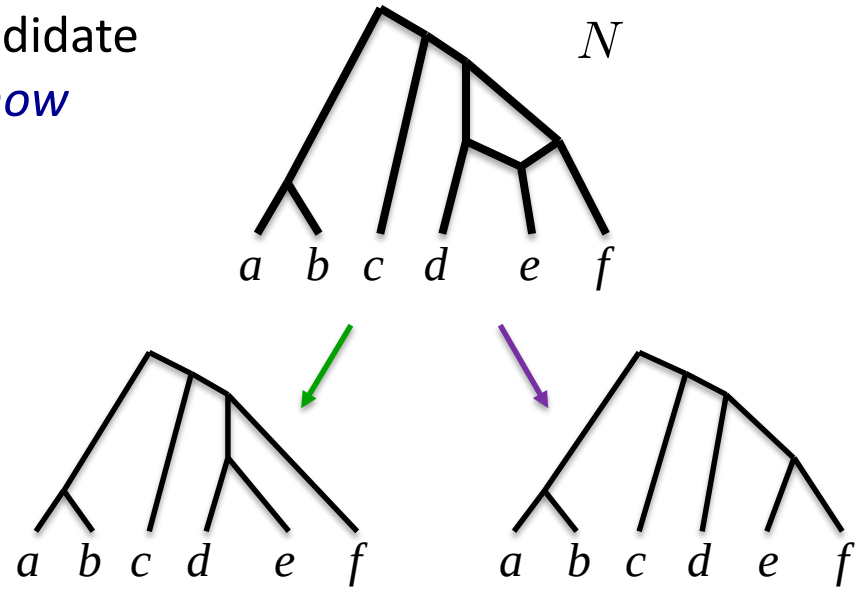
$A_1 \qquad A_2 \qquad \cdots \qquad A_m$

**Goal:**

Find $N$ that minimizes $\mathbf{Pr}(A_1, A_2, \ldots, A_m|N) = \prod_{i=1}^{m} \mathbf{Pr}(A_i|N) = \prod_{i=1}^{m} \left( \sum_{T \in \mathcal{T}(N)} \mathbf{Pr}(A_i|T)\mathbf{Pr}(T|N) \right)$

# Phylogenetic network inference

An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data:*




*Many possible formulations:*

**Data:**

Clusters of taxa: $\{a,b\}, \{d,e\}, \{d,e,f\}, \{a,b,c,d,e,f\}, \{e,f\}, \{c,d,e,f\}, \ldots$

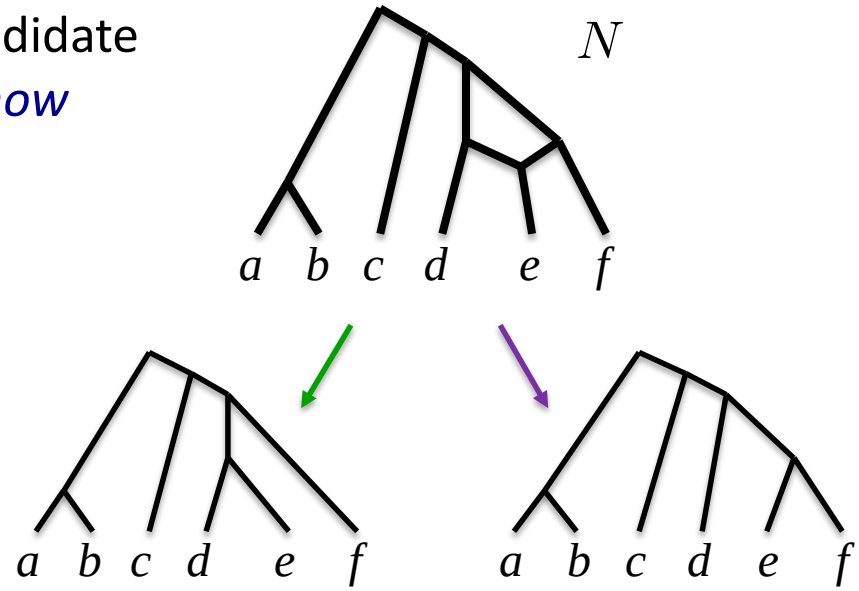
**Goal:**

Find $N$ that maximizes the number of input clusters that are `explained' by one of the trees displayed by $N$

subject to constraints on the complexity of $N$

# Phylogenetic network inference

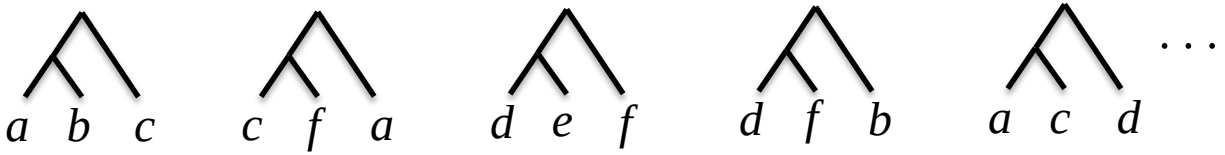
An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data:*




*Many possible formulations:*

**Data:**

Trees with 3 taxa:
(inferred from other data)



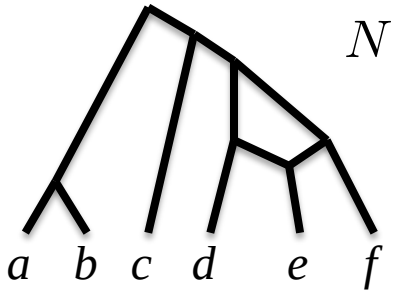

**Goal:**

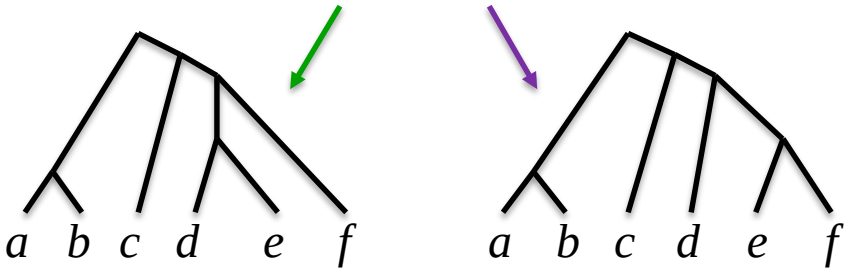
Find $N$ that maximizes the number of input trees that are `consistent' with one of the trees displayed by $N$

subject to constraints on the complexity of $N$

# Phylogenetic network inference

An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data:*

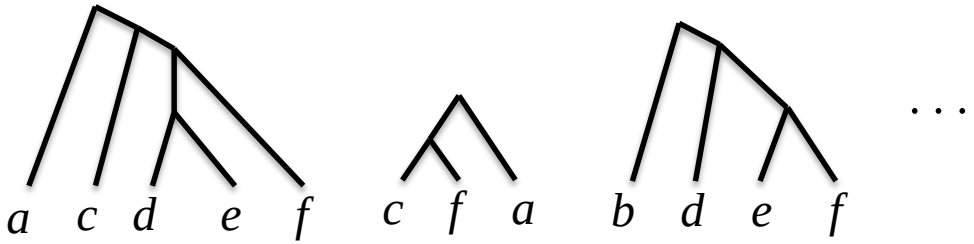



*Many possible formulations:*

**Data:**

Any trees on the same taxa:
(inferred from other data)




**Goal:**

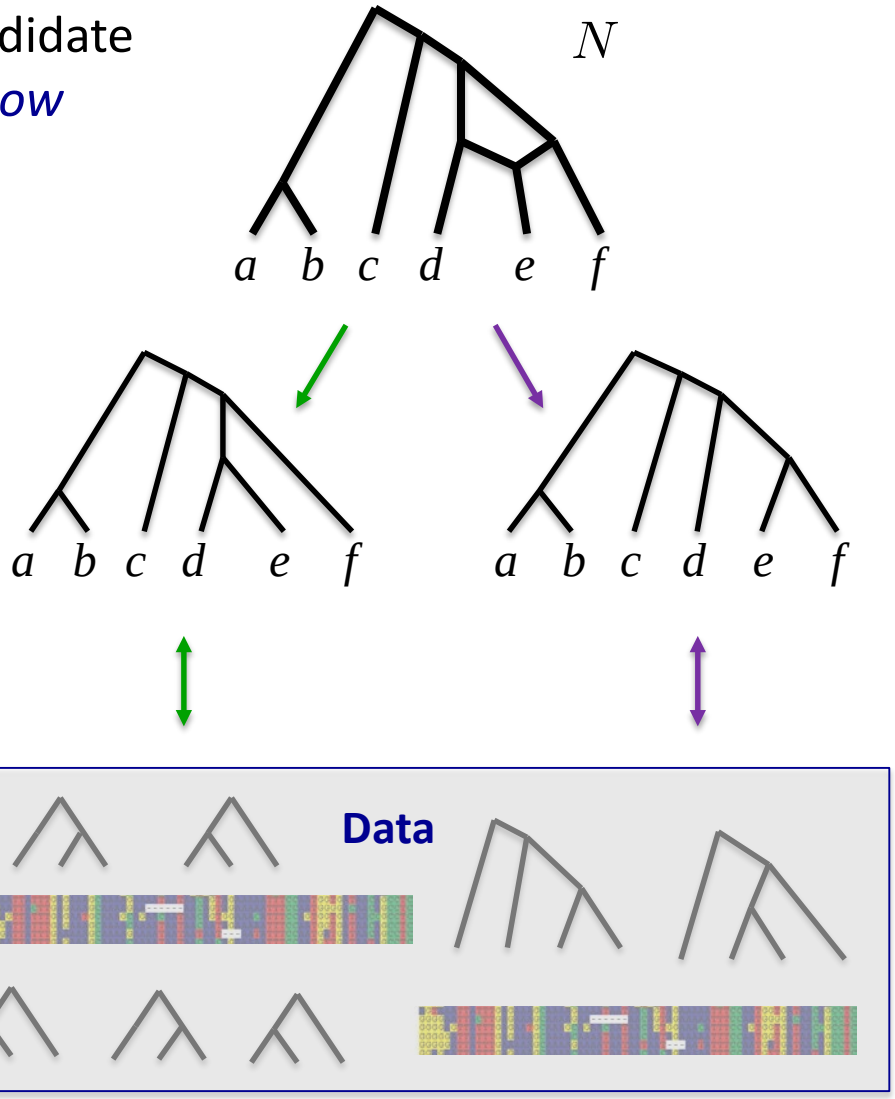
Find $N$ that maximizes the number of input trees that are `consistent' with one of the trees displayed by $N$

subject to constraints on the complexity of $N$

# Phylogenetic network inference

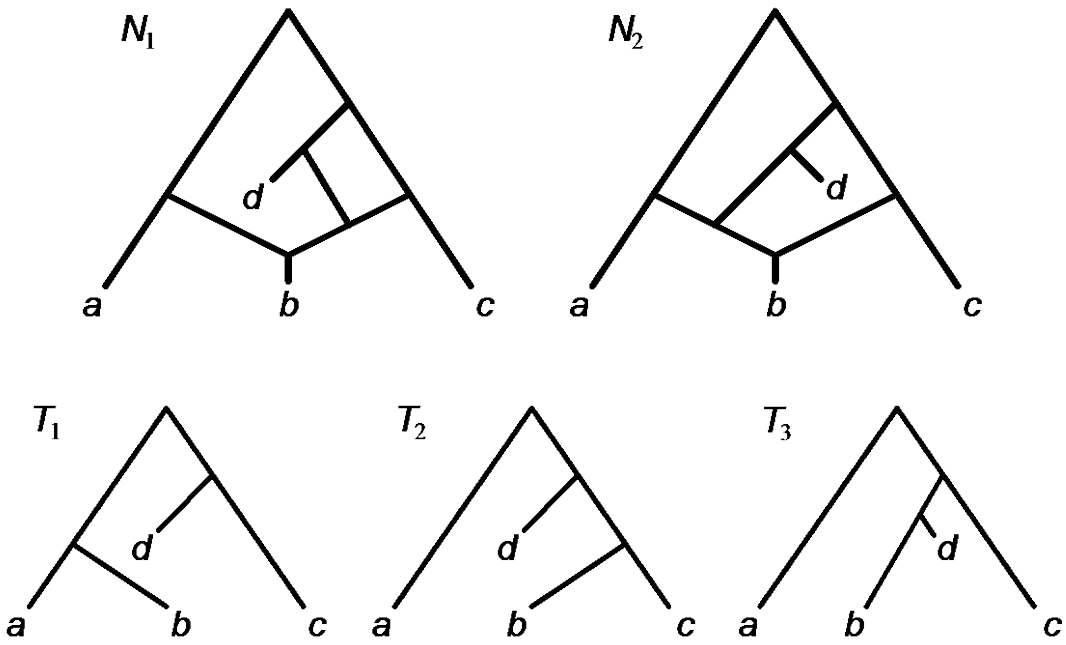
An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data:*

*Many possible formulations…*



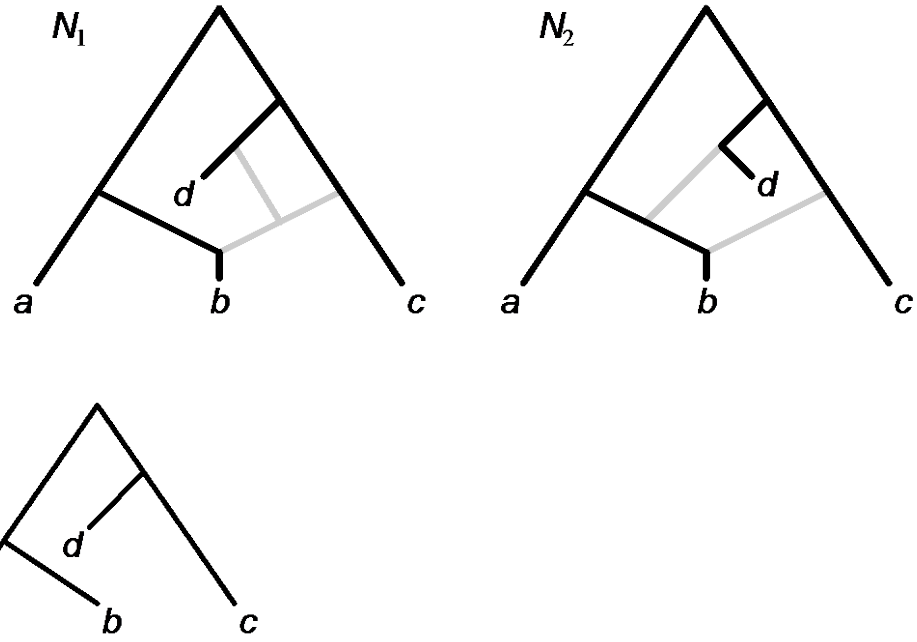
# Different networks can display the same trees

Some networks display exactly the same trees:



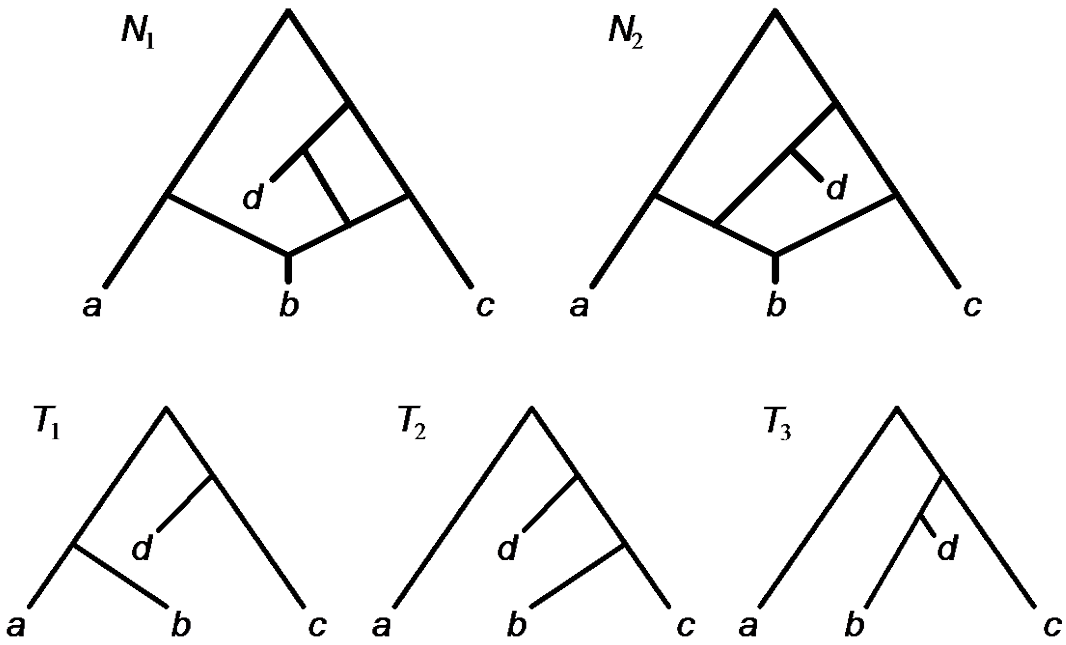
# Different networks can display the same trees

Some networks display exactly the same trees:



# Different networks can display the same trees

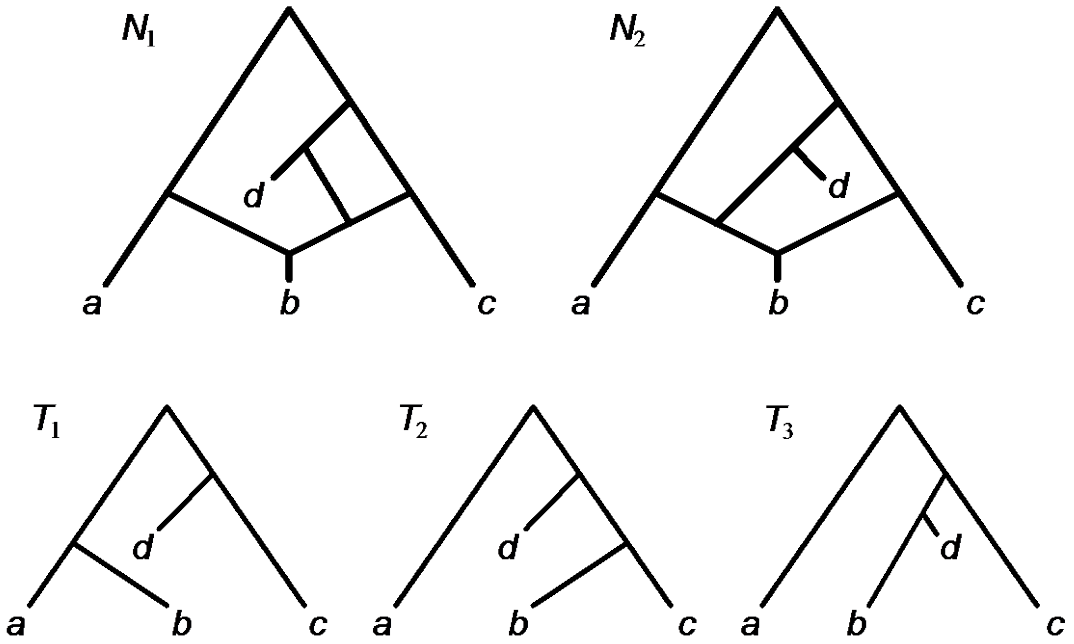
Some networks display exactly the same trees:



# Different networks can display the same trees

Some networks display exactly the same trees:

Because $N_1$ and $N_2$ display the same trees, they are equally good to any of the inference methods we saw – *no matter the input data*

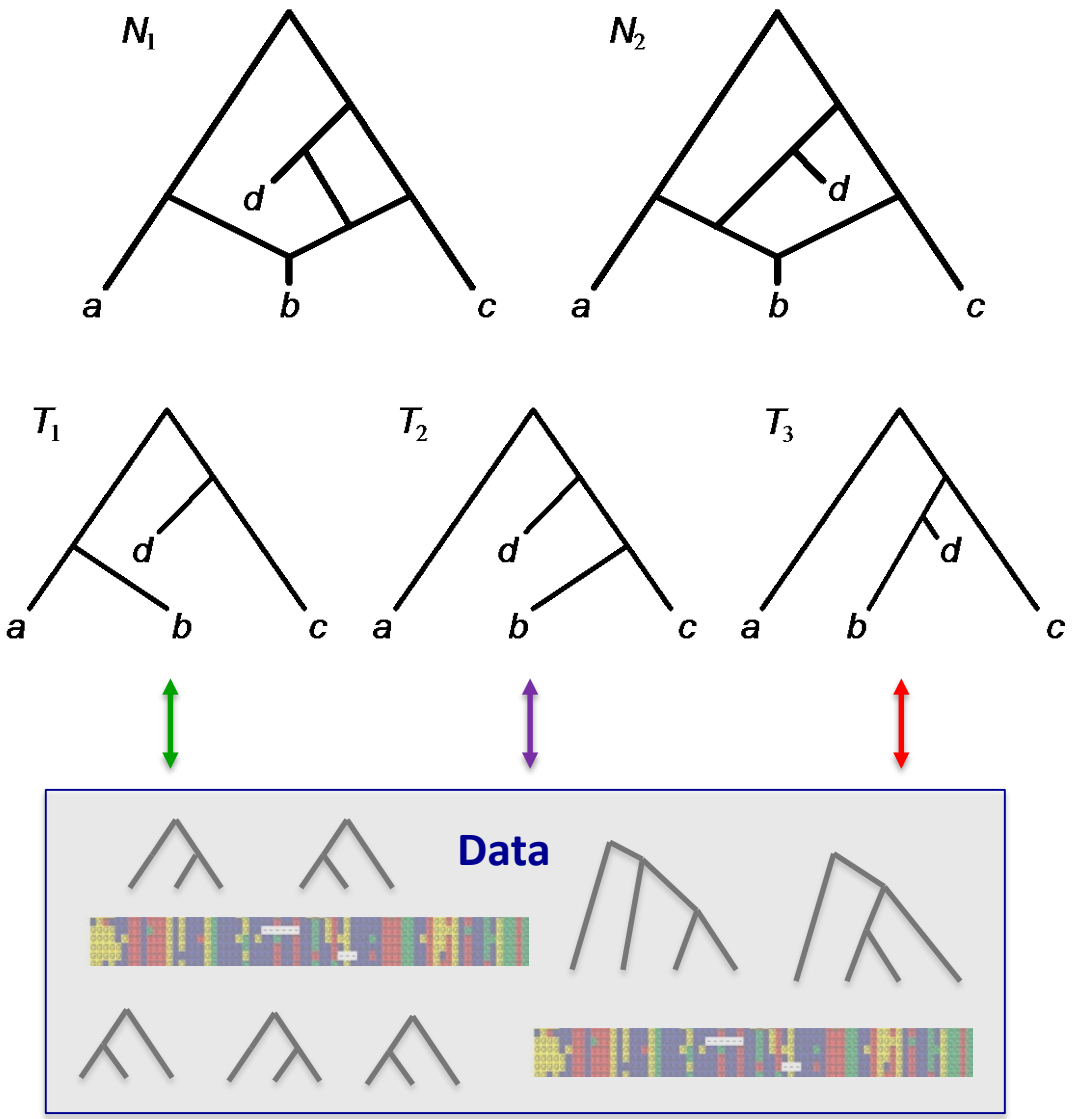


# Different networks can display the same trees

Some networks display exactly the same trees:

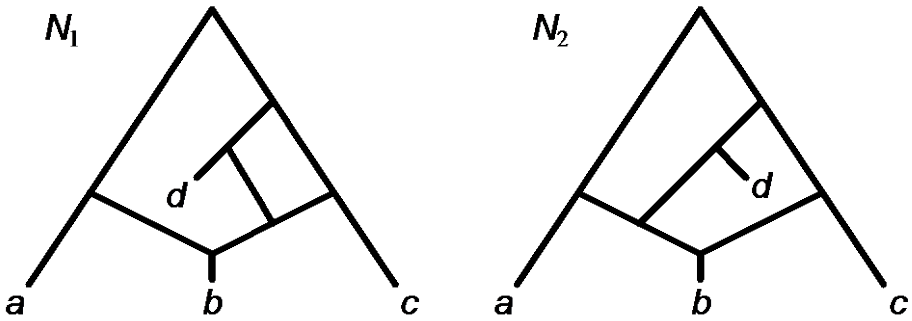
Because $N_1$ and $N_2$ display the same trees, they are equally good to any of the inference methods we saw – *no matter the input data*

(Recall that a network is evaluated *on the basis of how well the trees it displays fit the data*)



# Different networks can display the same trees

Some networks display exactly the same trees:

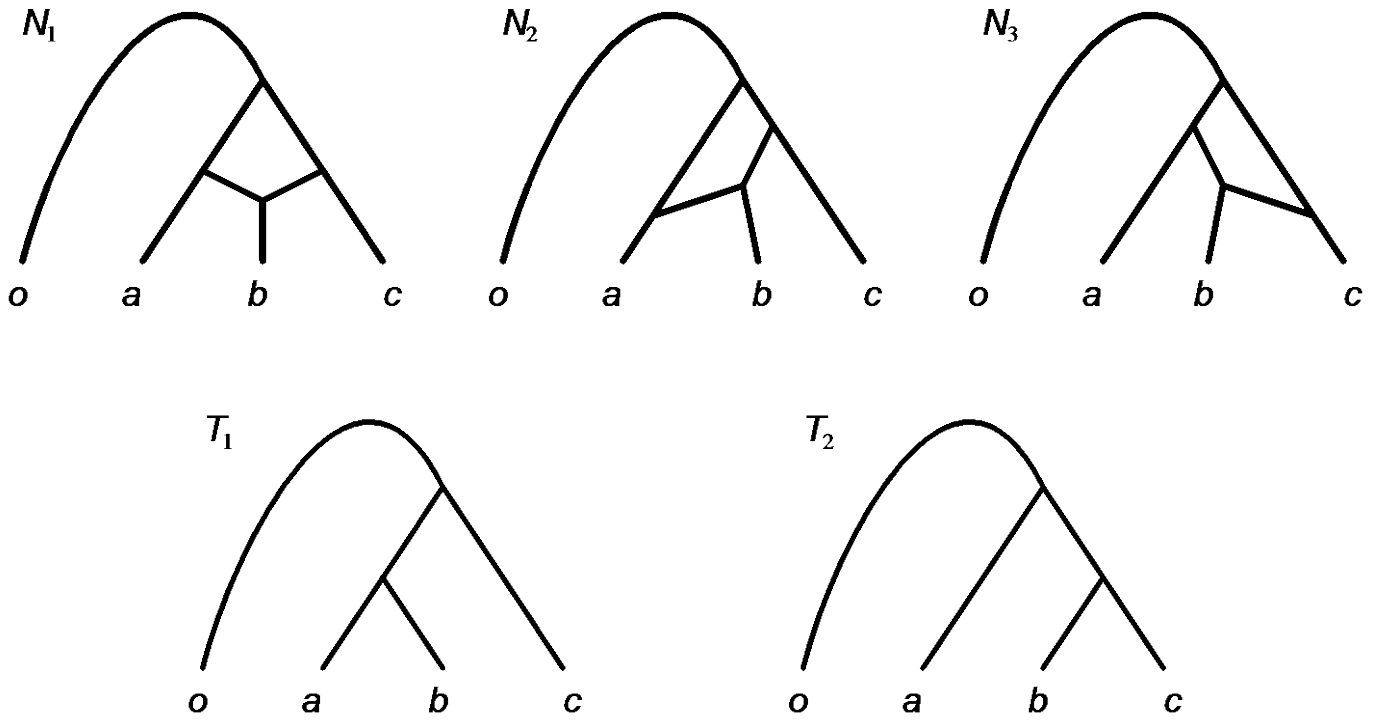



Because $N_1$ and $N_2$ display the same trees, they are equally good to any of the inference methods we saw – *no matter the input data*

UNIDENTIFIABILITY

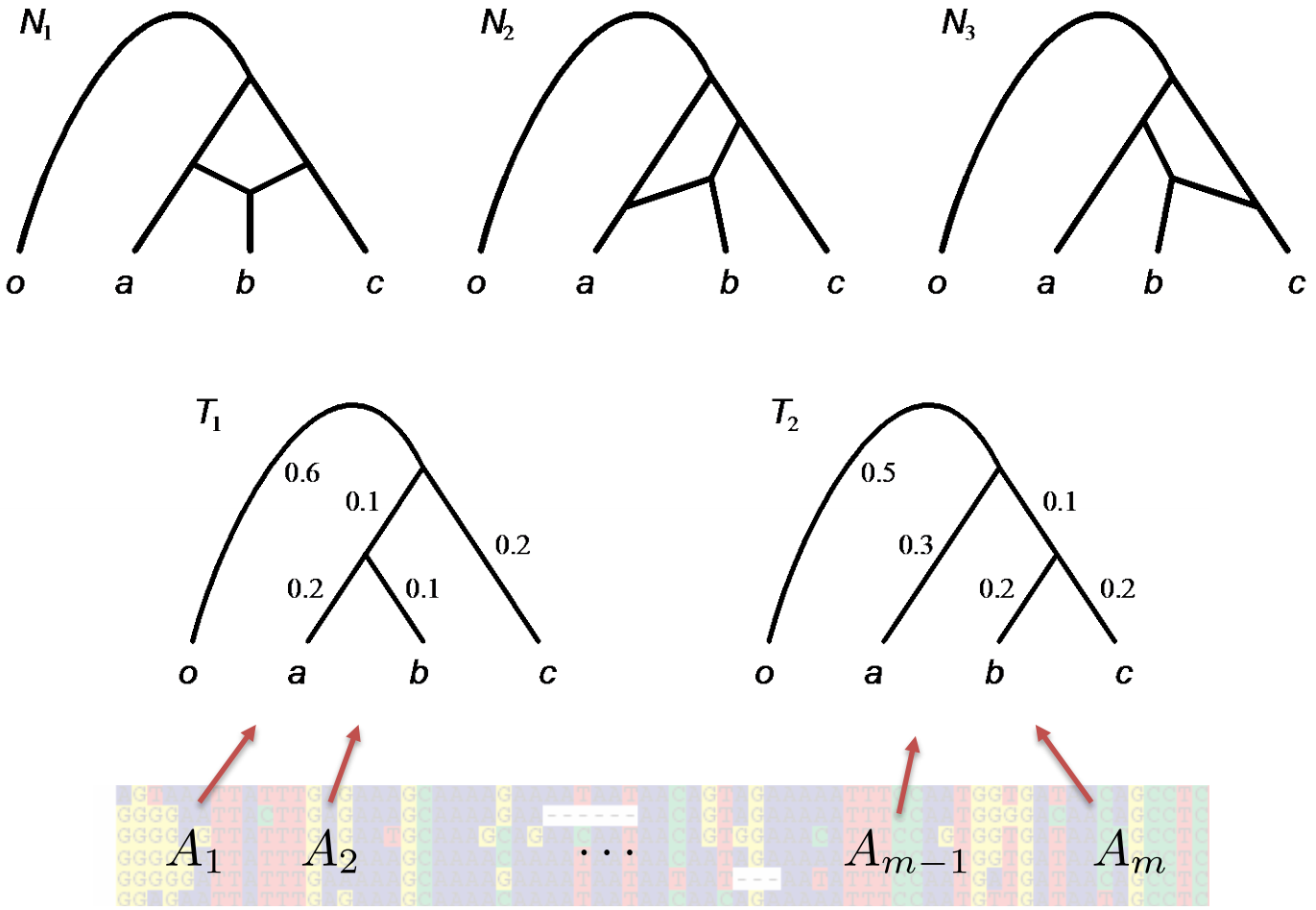
# Branch lengths are informative

Branch lengths can be used to distinguish between otherwise indistinguishable scenarios:



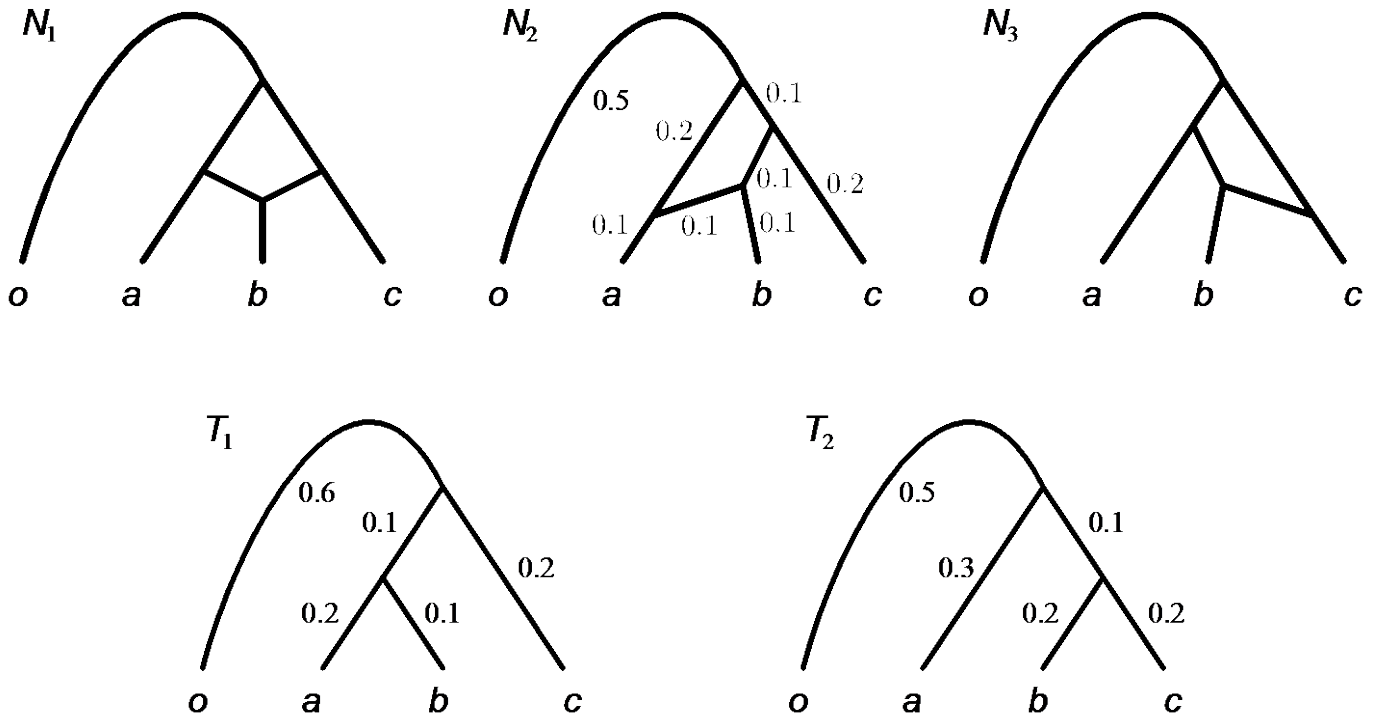
# Branch lengths are informative

Branch lengths can be used to distinguish between otherwise indistinguishable scenarios:



# Branch lengths are informative

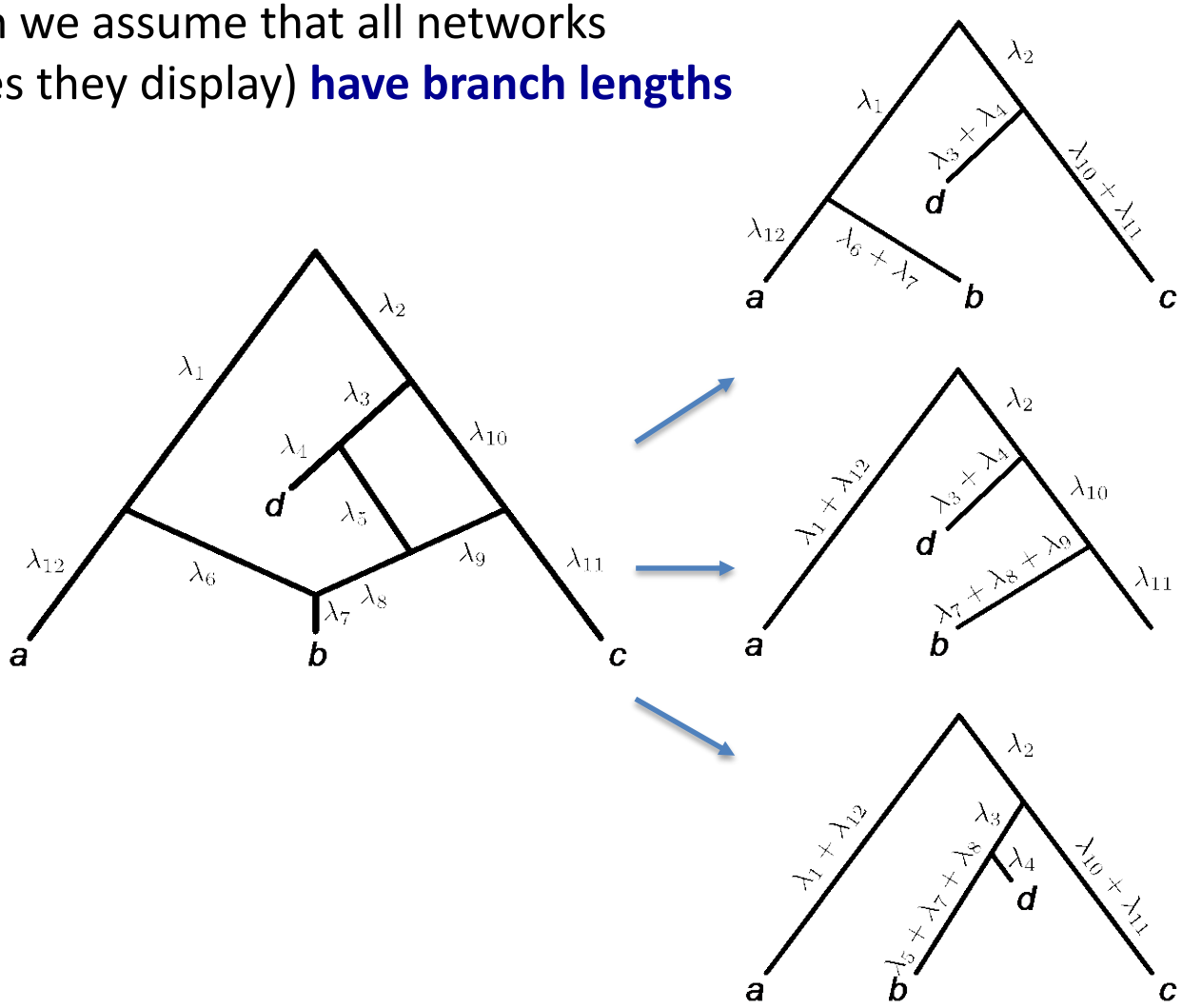
Branch lengths can be used to distinguish between otherwise indistinguishable scenarios:




$N_2$ is the only network to which we can assign branch lengths so that it displays $T_1$ and $T_2$

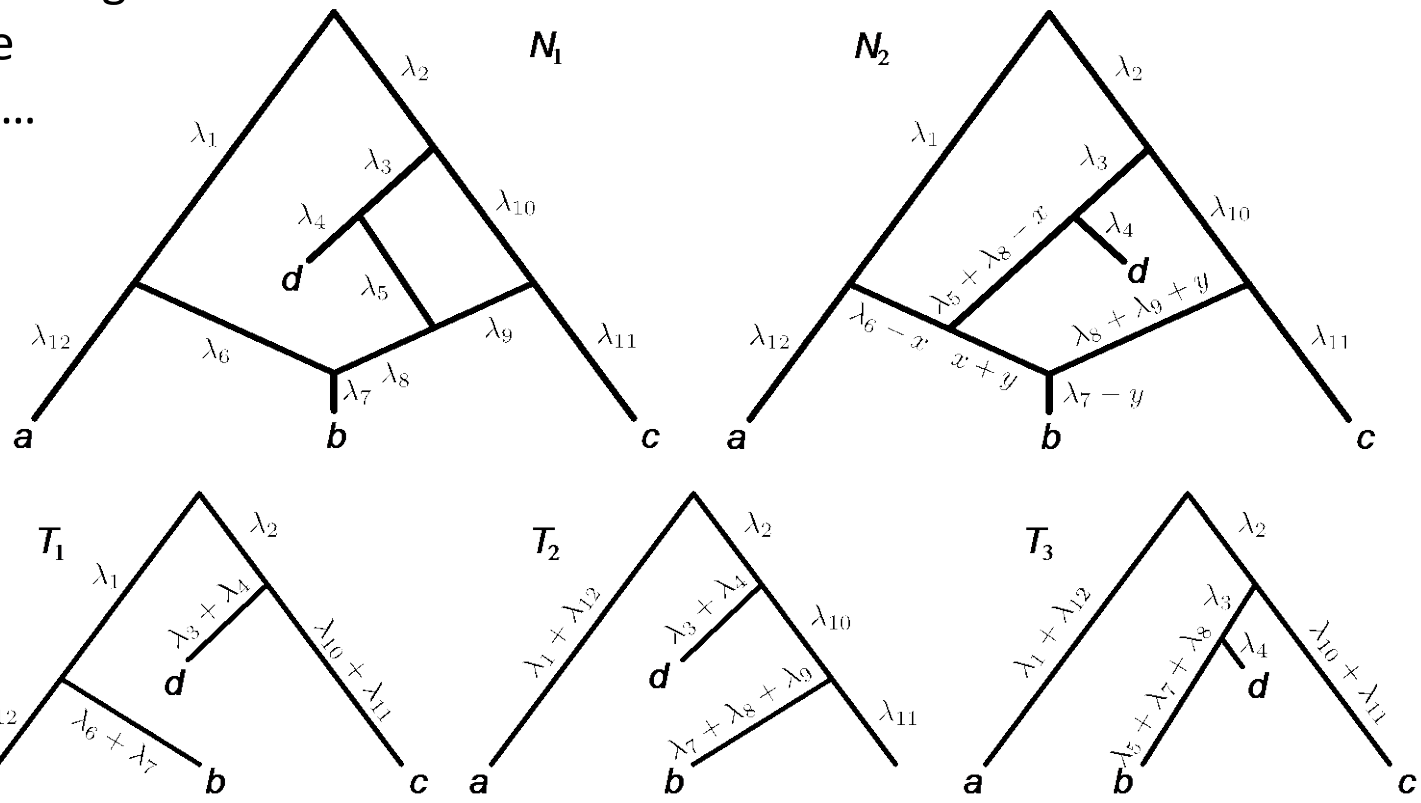
# Branch lengths are informative

From now on we assume that all networks
(and the trees they display) **have branch lengths**



# Indistinguishable networks

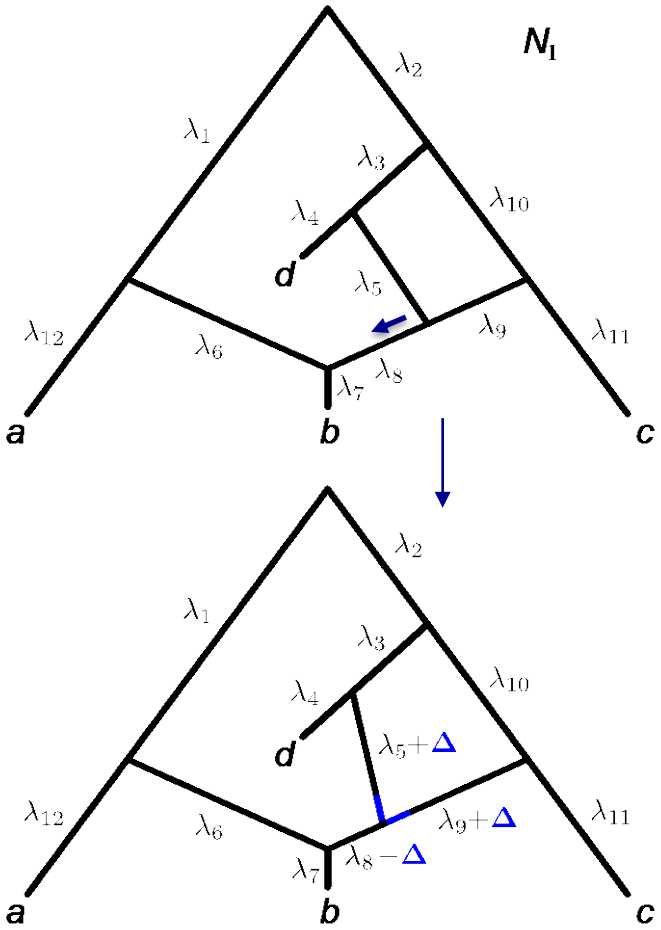
However, branch lengths do not eliminate unidentifiability…




$N_1$ and $N_2$ display the same trees (i.e. including branch lengths) and are thus *indistinguishable* even to methods accounting for lengths

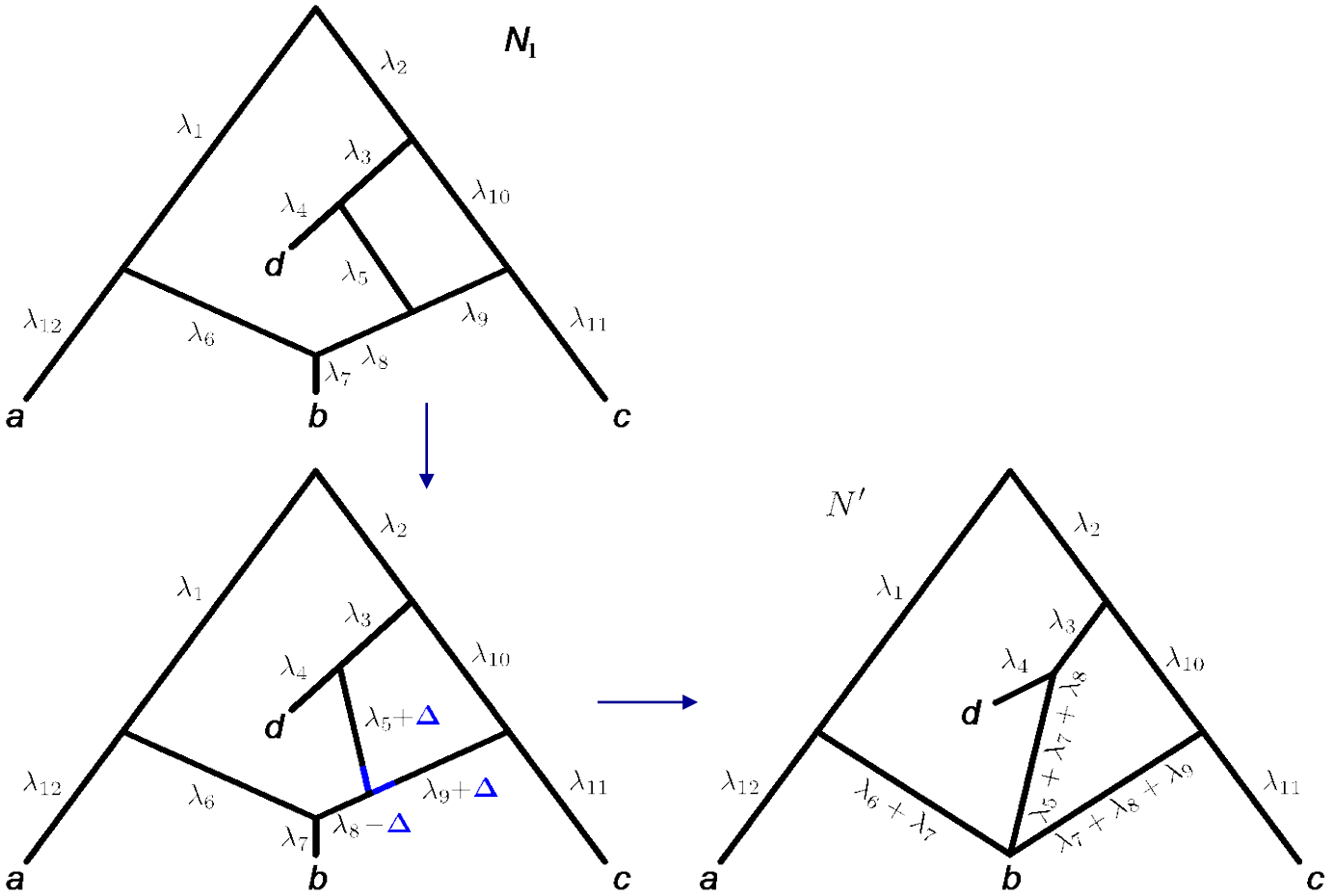
# Unzipping a network

Key observation: we can move reticulations up or down (until they hit a speciation node) and the trees displayed by a network remain the same:



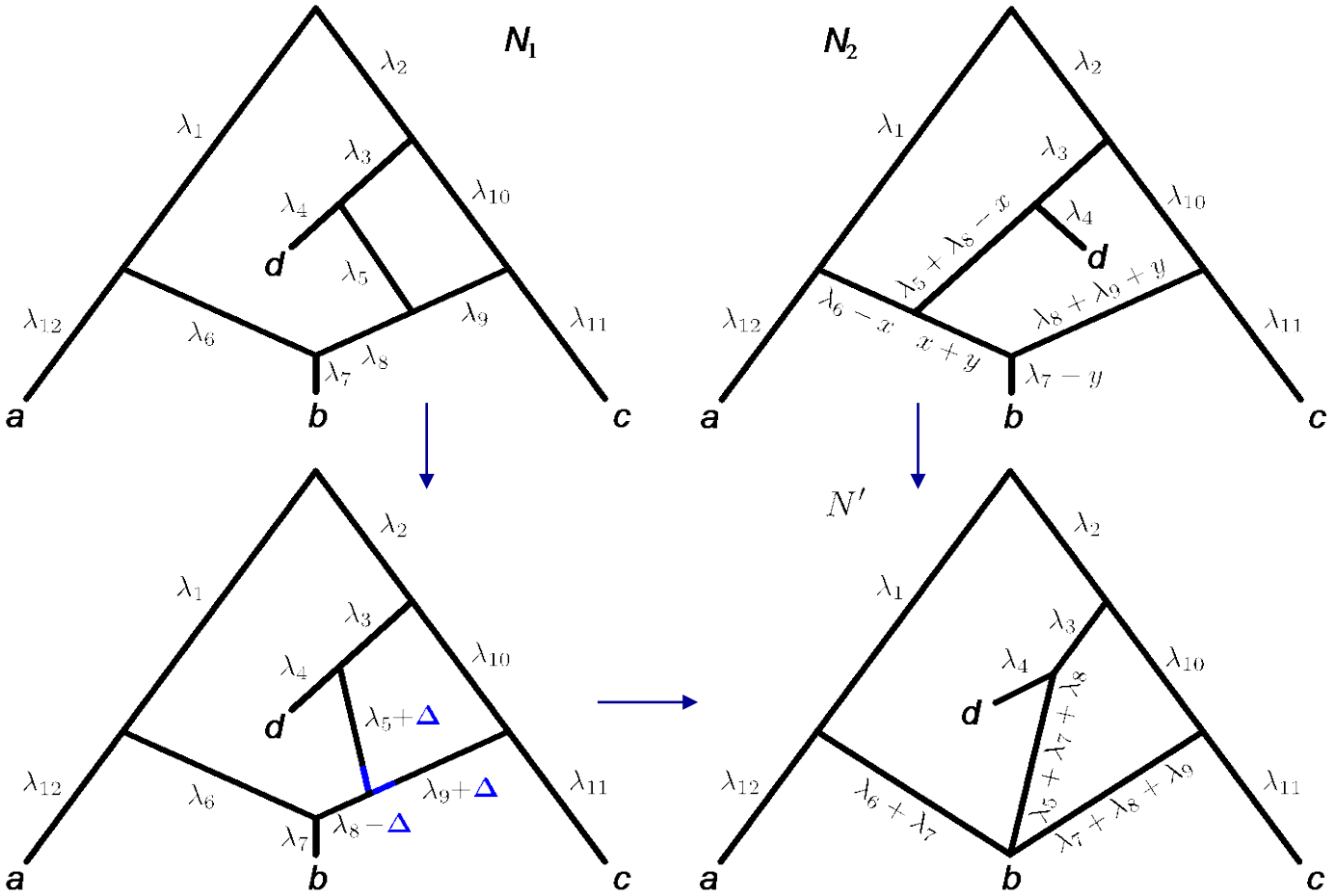
# Unzipping a network

Key observation: we can move reticulations up or down (until they hit a speciation node) and the trees displayed by a network remain the same:



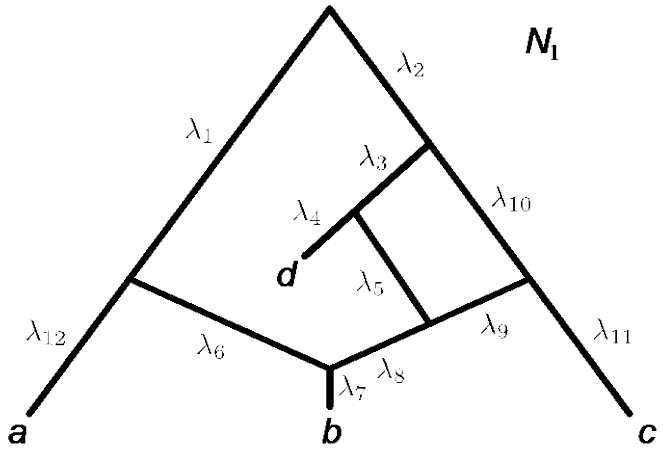
# Unzipping a network

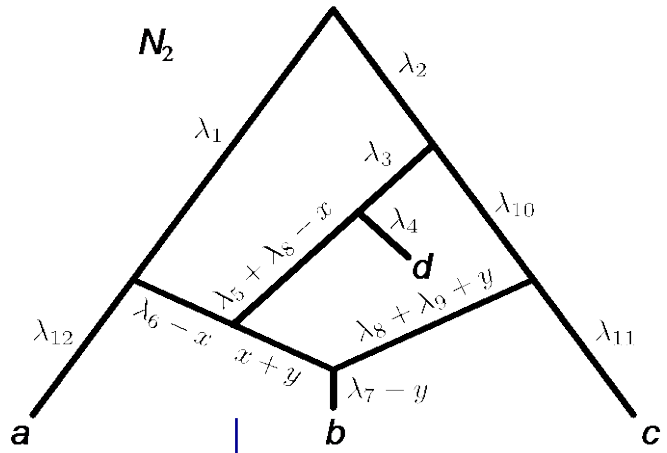
Key observation: we can move reticulations up or down (until they hit a speciation node) and the trees displayed by a network remain the same:



# Unzipping a network

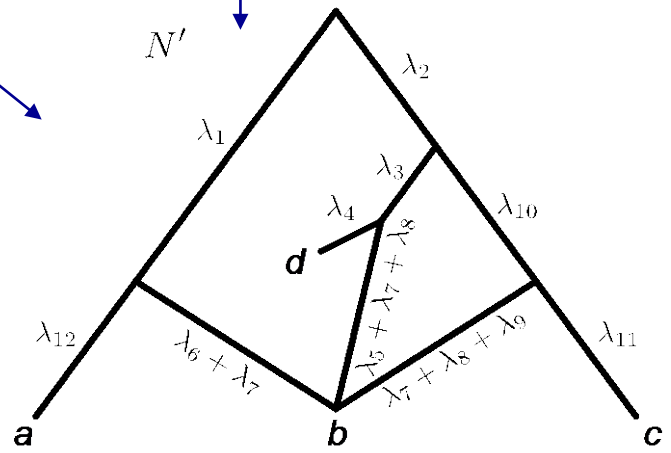
Key observation: we can move reticulations up or down (until they hit a speciation node) and the trees displayed by a network remain the same:



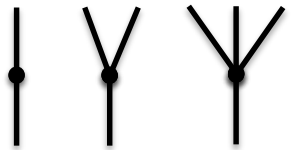

By moving reticulations always down, $N_1$ and $N_2$ both end up becoming the same network.

True in general: indistinguishable networks always transform into the same network.
We call this the *canonical form* of $N_1$ and $N_2$.



# Take home message (1) for the mathematician

- $N_1$ and $N_2$ are *indistinguishable* if they display the same trees (with branch lengths)
- A *funnel* is a node with indegree > 0 and outdegree = 1:

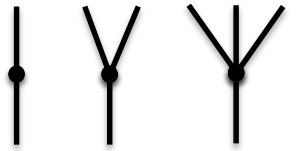


...

- $N*$ is the *canonical form* of $N$ if:
  $N*$ is indistinguishable from $N$ and
  $N*$ has no funnel

# Take home message (1) for the mathematician

- $N_1$ and $N_2$ are *indistinguishable* if they display the same trees (with branch lengths)
- A *funnel* is a node with indegree > 0 and outdegree = 1:

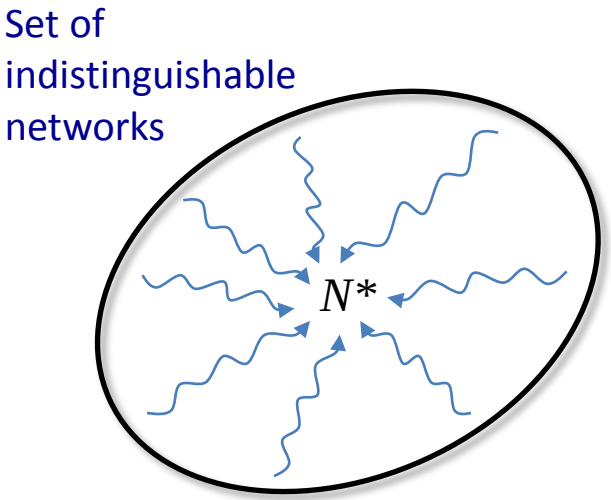


...

- $N*$ is the *canonical form* of $N$ if:
  $N*$ is indistinguishable from $N$ and
  $N*$ has no funnel

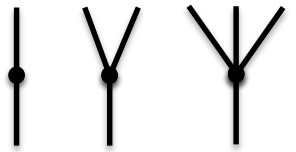
**Theorem**

- Every network has a canonical form
- The canonical form of a network is unique
  (under mild assumptions – 'U.m.a.')



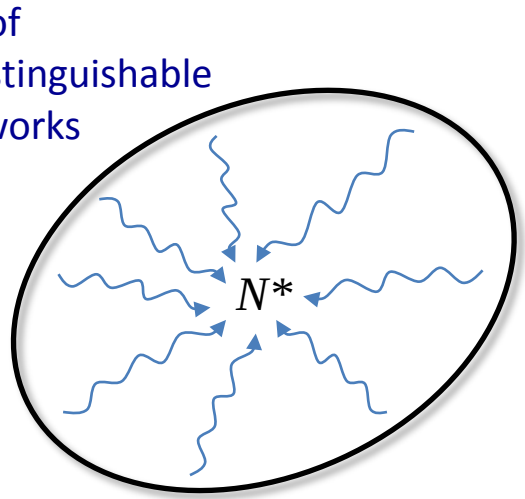
# Take home message (1) for the mathematician

- $N_1$ and $N_2$ are *indistinguishable* if they display the same trees (with branch lengths)
- A *funnel* is a node with indegree > 0 and outdegree = 1:



...

- $N*$ is the *canonical form* of $N$ if:
  $N*$ is indistinguishable from $N$ and
  $N*$ has no funnel




**Theorem**

- Every network has a canonical form
- The canonical form of a network is unique
  (under mild assumptions – 'U.m.a.')

**Corollary 1**

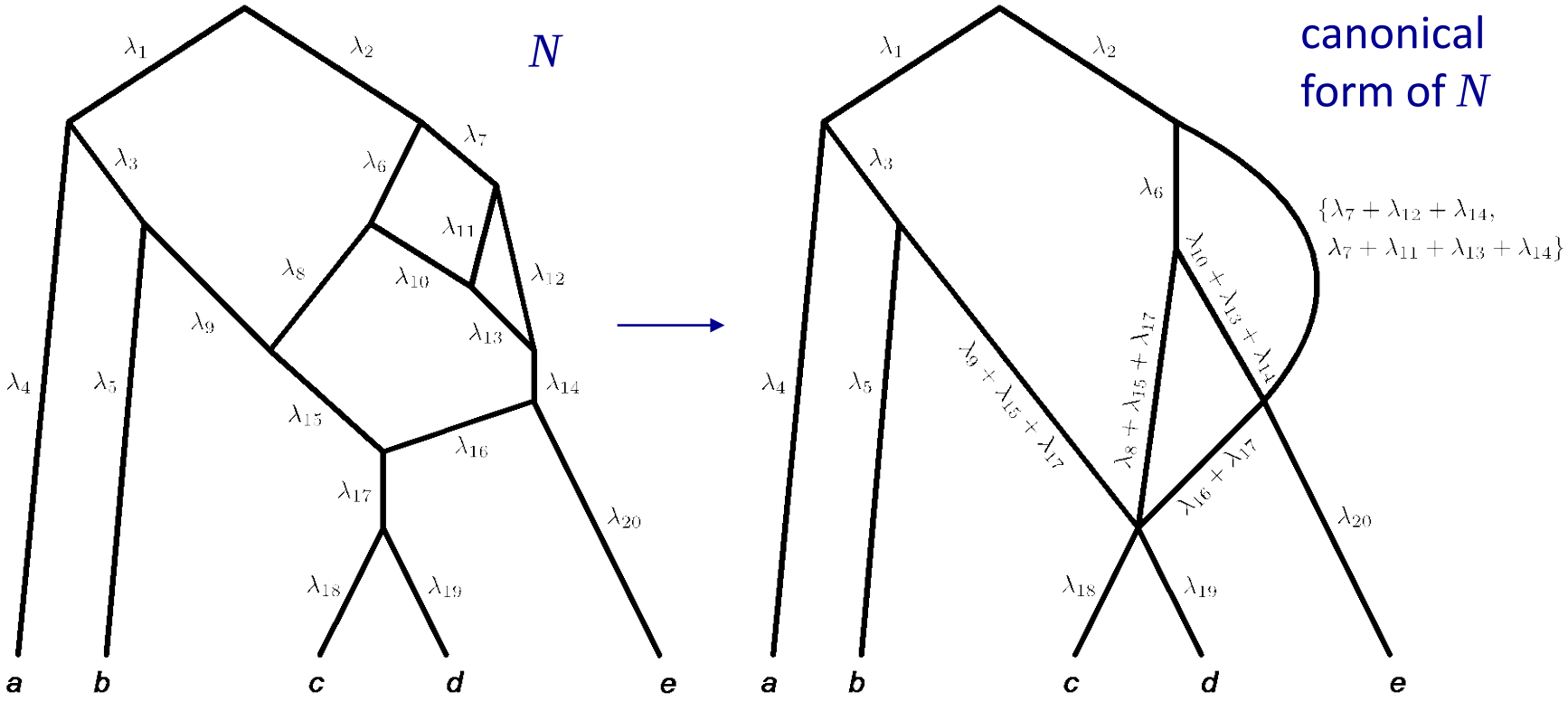
U.m.a., $N_1$ and $N_2$ are indistinguishable iff they have the same canonical form

**Corollary 2**

U.m.a., a network in canonical form is uniquely determined by the trees it displays

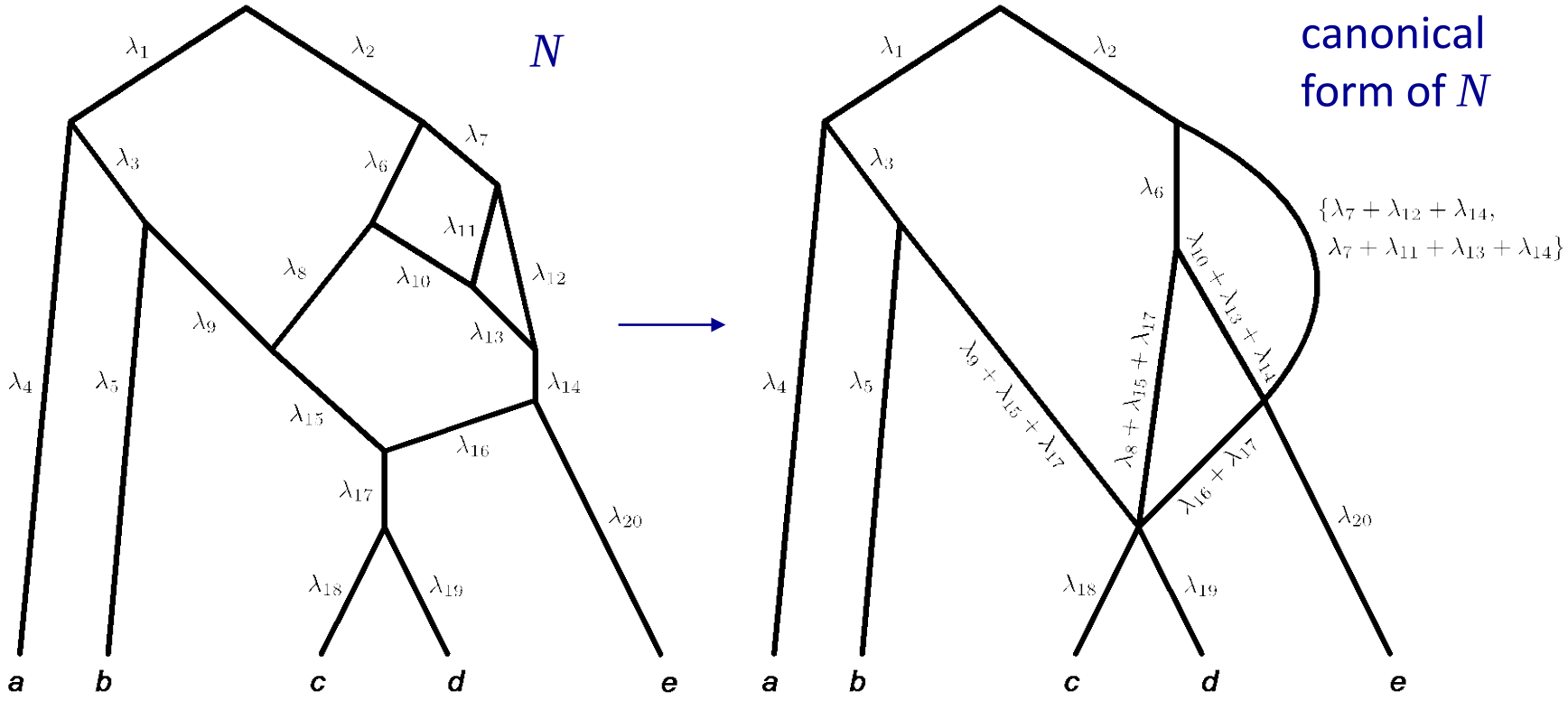
# Take home message (2) for the biologist

The canonical form of a network $N$ is a simplified, but 'equivalent`, version of $N$ that excludes all unrecoverable aspects of $N$. For example:



# Take home message (2) for the biologist

The canonical form of a network $N$ is a simplified, but 'equivalent`, version of $N$ that excludes all unrecoverable aspects of $N$. For example:




If $N$ is reconstructed by an inference method, then even assuming perfect data, the true phylogenetic network is just one of the many that are indistinguishable from $N$ … the canonical form is representative of all of them.

# Take home message (3) for the computational biologist

Network inference methods should only attempt to reconstruct what they can uniquely identify: canonical forms

**Theorem**

- Every network has a canonical form
- The canonical form of a network is unique (under mild assumptions)

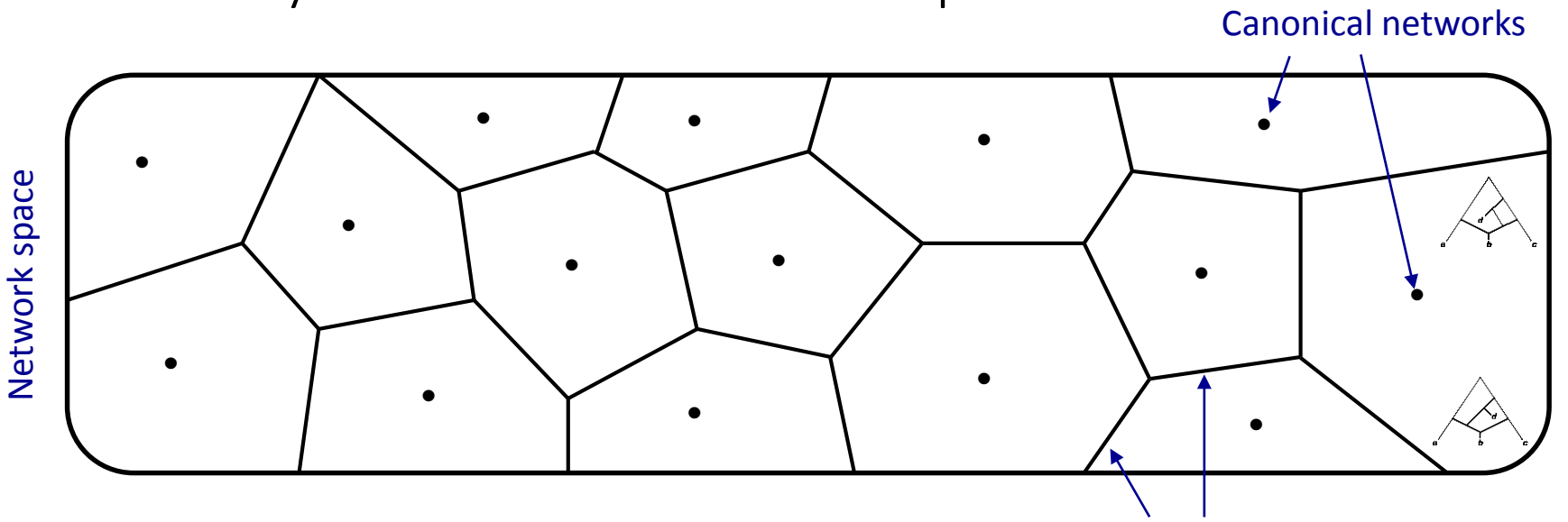
# Take home message (3) for the computational biologist

Network inference methods should only attempt to reconstruct what they can uniquely identify: canonical forms

**Theorem**

- Every network has a canonical form
- The canonical form of a network is unique (under mild assumptions)

Instead of searching (or directly constructing) within network space, one should carry out the inference in a reduced space:




- Should reduce computation time
- Partially address the problem of multiple optimal networks

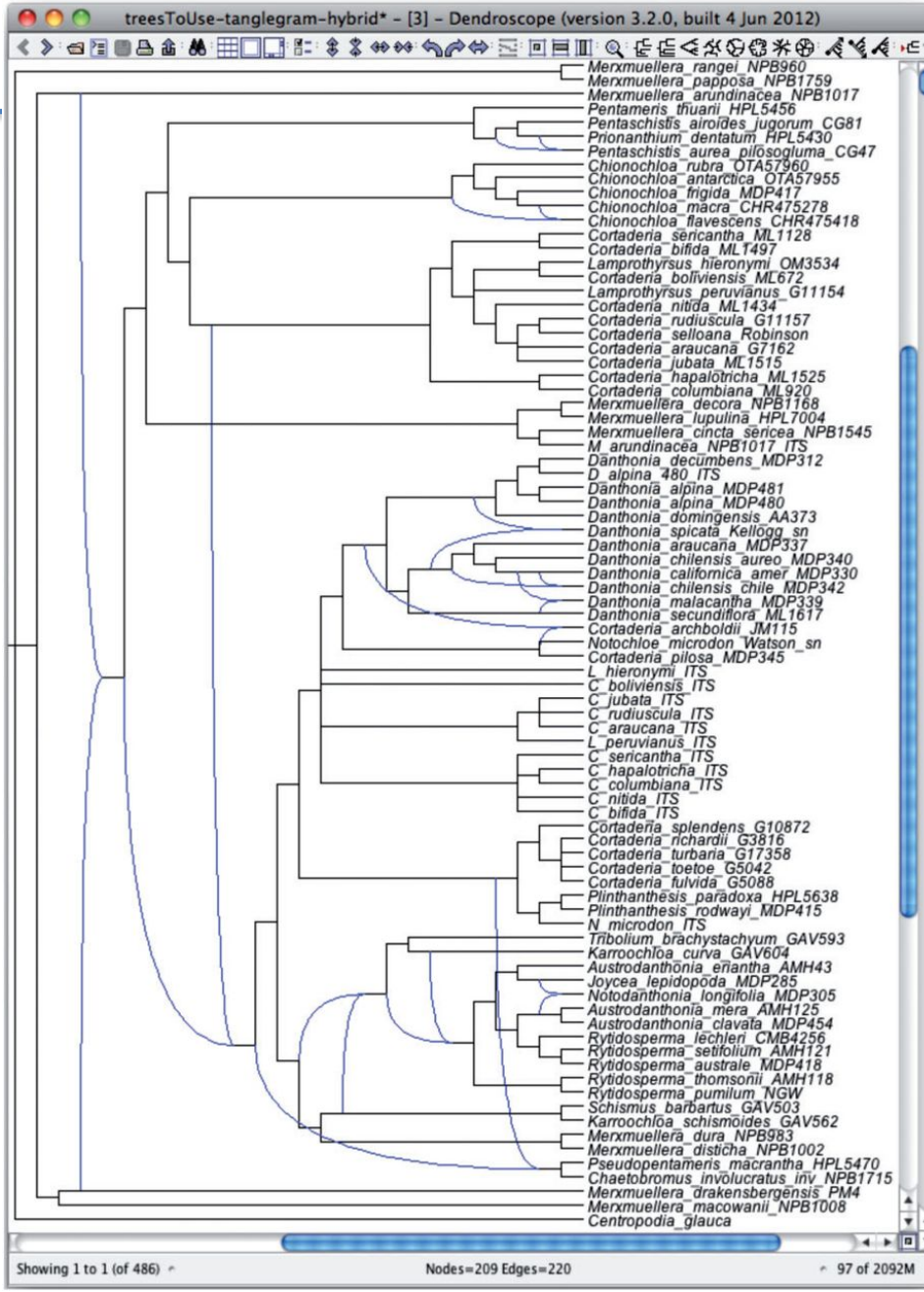
# Multiple optimal networks

Inferring networks in canonical form should partially address the problem of multiple optima :

Huson and Scornavacca. *Syst Biol* 2012:

*A minimum hybridization network computed by Dendroscope 3 […] It is one of **486** networks calculated by the program.*

It is not hard to see that some of these 486 networks are simply indistinguishable



# Canonical forms for the mathematician, again

The existence is proven with a simple reduction algorithm

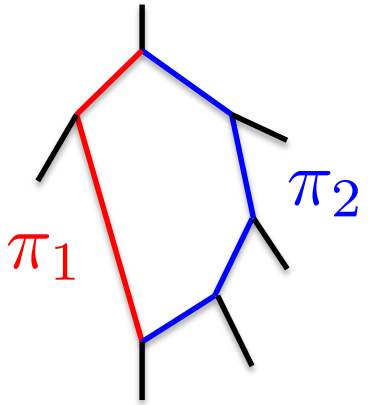
Uniqueness is much harder to prove and relies on the network satisfying the following property:

**Theorem**

- Every network has a canonical form
- The canonical form of a network is unique (under mild assumptions)

*No pair of distinct paths having the same endpoints have the same length*

$$\sum_{e \in \pi_1} \lambda_e \neq \sum_{e \in \pi_2} \lambda_e$$



# Canonical forms for the mathematician, again

The existence is proven with a simple reduction algorithm

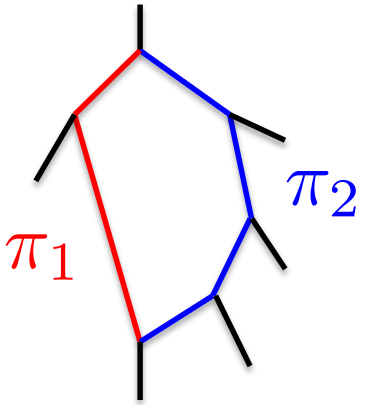
Uniqueness is much harder to prove and relies on the network satisfying the following property:

**Theorem**

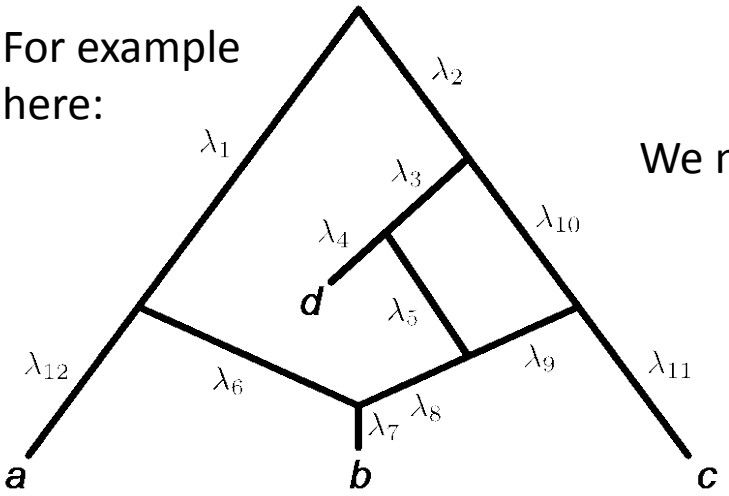
- Every network has a canonical form
- The canonical form of a network is unique (under mild assumptions)

*No pair of distinct paths having the same endpoints have the same length*

$$\sum_{e\in\pi_1} \lambda_e \neq \sum_{e\in\pi_2} \lambda_e$$




For example here:




We must impose:

$$\lambda_1 + \lambda_6 \neq \lambda_2 + \lambda_3 + \lambda_5 + \lambda_8$$
$$\lambda_1 + \lambda_6 \neq \lambda_2 + \lambda_{10} + \lambda_9 + \lambda_8$$
$$\lambda_2 + \lambda_3 + \lambda_5 + \lambda_8 \neq \lambda_2 + \lambda_{10} + \lambda_9 + \lambda_8$$

which happens with probability 1

Thank you for your attention!