Methods to reconstruct phylogenetic networks accounting for ILS

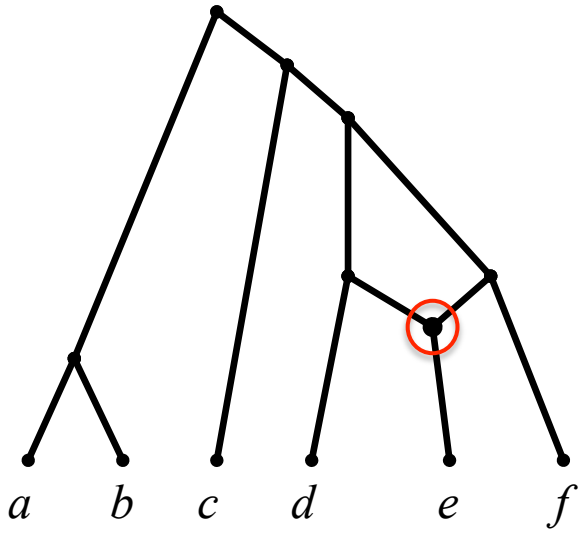
Céline Scornavacca
some slides have been kindly provided by Fabio Pardi

ISE-M, Equipe Phylogénie & Evolution Moléculaires
Montpellier, France

February the 2nd, 2017

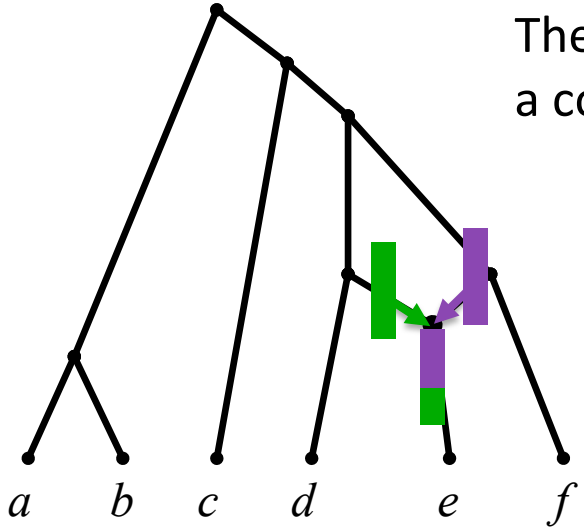
Trees displayed by a network

In a phylogenetic network, a reticulate event is represented as a reticulation, where branches converge to give rise to a new lineage:



Trees displayed by a network

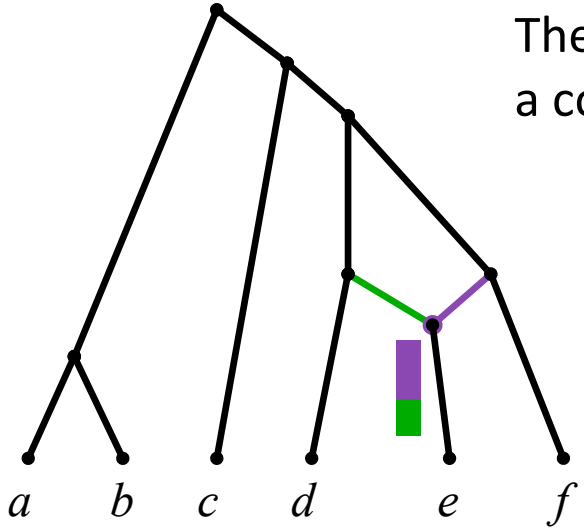
In a phylogenetic network, a reticulate event is represented as a reticulation, where branches converge to give rise to a new lineage:



The genome(s) at the start of the new lineage is (are) a composition of those of the parent lineages.

Trees displayed by a network

In a phylogenetic network, a reticulate event is represented as a reticulation, where branches converge to give rise to a new lineage:

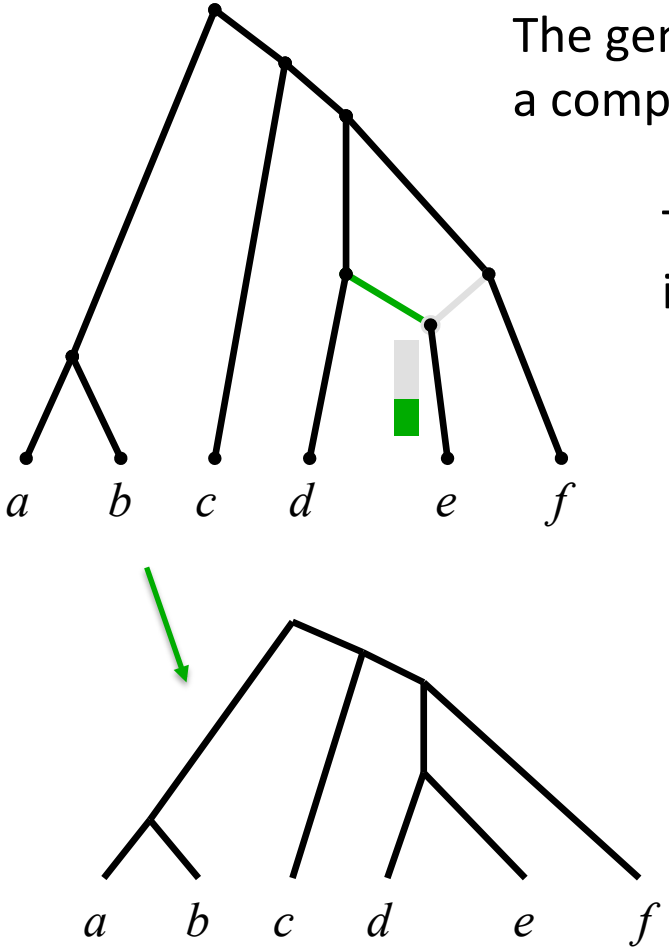


The genome(s) at the start of the new lineage is (are) a composition of those of the parent lineages.

The evolution of each part independently inherited is described by a *gene tree*

Trees displayed by a network

In a phylogenetic network, a reticulate event is represented as a reticulation, where branches converge to give rise to a new lineage:

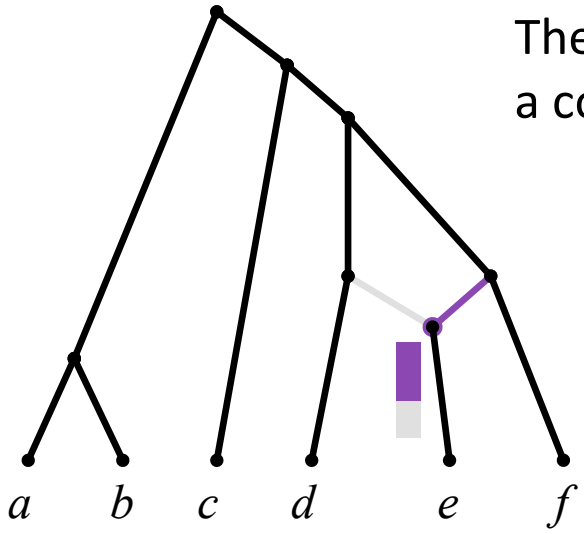


The genome(s) at the start of the new lineage is (are) a composition of those of the parent lineages.

The evolution of each part independently inherited is described by a *gene tree*

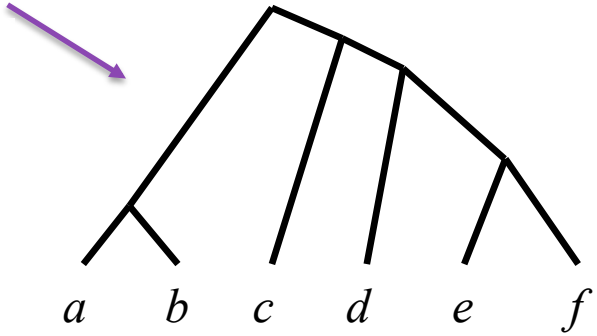
Trees displayed by a network

In a phylogenetic network, a reticulate event is represented as a reticulation, where branches converge to give rise to a new lineage:



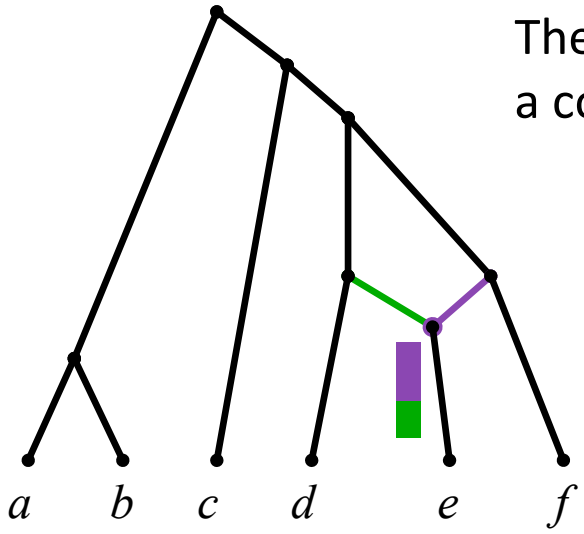
The genome(s) at the start of the new lineage is (are) a composition of those of the parent lineages.

The evolution of each part independently inherited is described by a *gene tree*



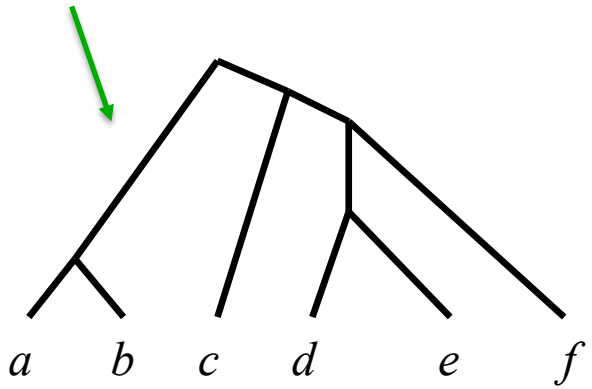
Trees displayed by a network

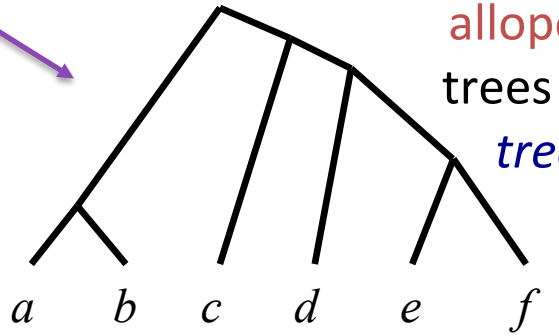
In a phylogenetic network, a reticulate event is represented as a reticulation, where branches converge to give rise to a new lineage:



The genome(s) at the start of the new lineage is (are) a composition of those of the parent lineages.

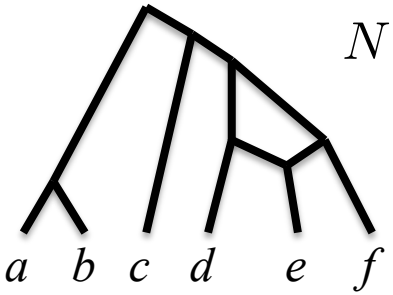
The evolution of each part independently inherited is described by a *gene tree*



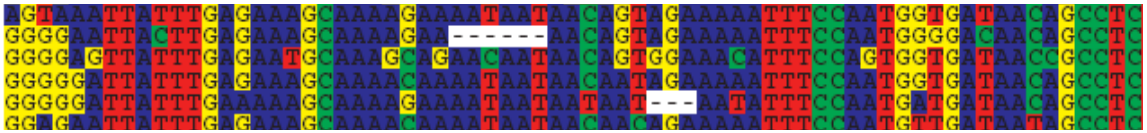


In the absence of deep coalescence and allopolyploidy, the gene trees are a subset of the *trees displayed* by the network

Phylogenetic network inference

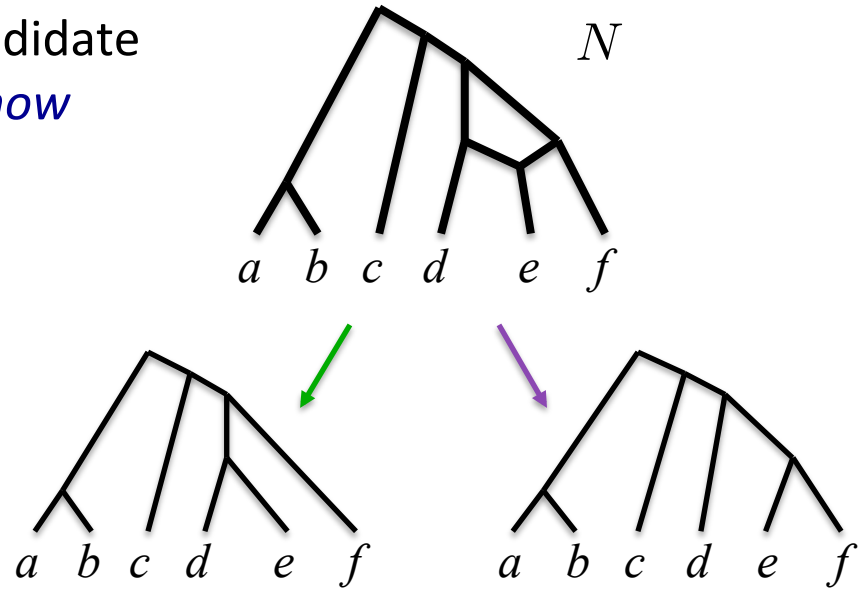


?



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)

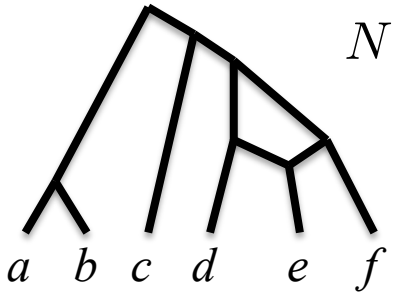
Goal:

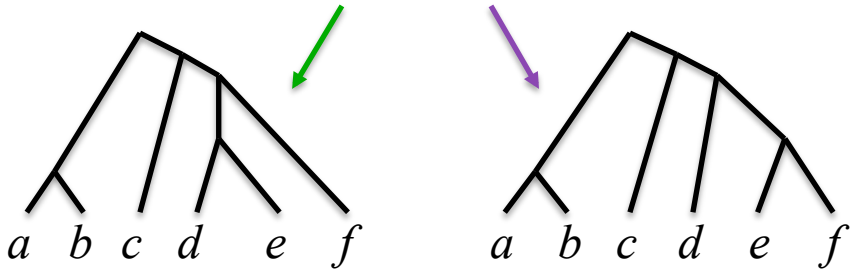
Find the network N with the lower hybridization number such that the triplets are `consistent' with one of the trees displayed by N

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)

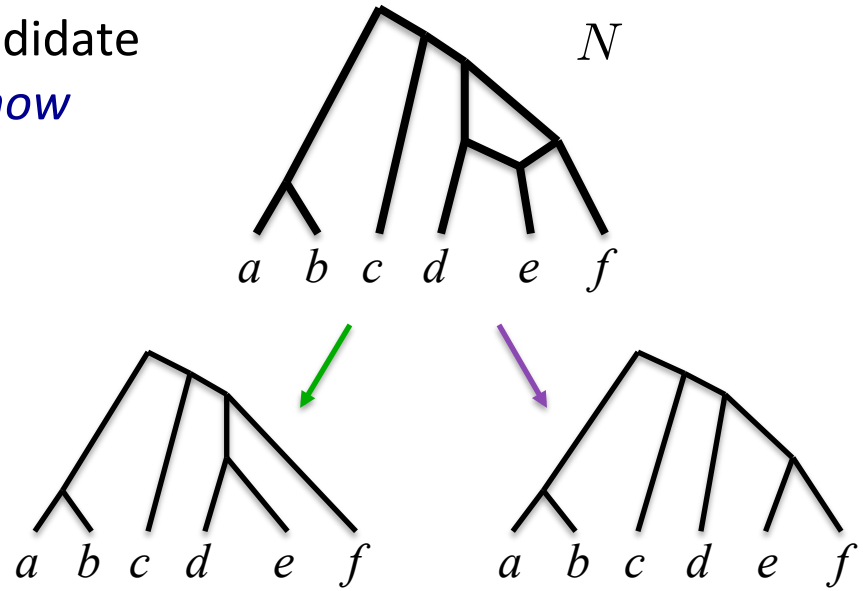
Goal:

Find the network N with the lower hybridization number such that the input trees 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:

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

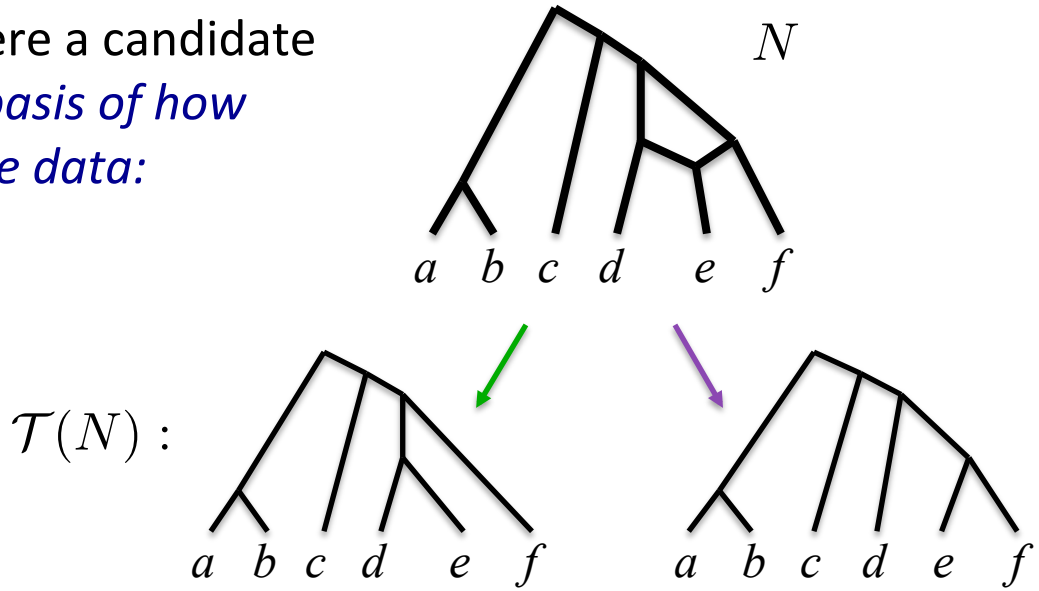
Goal:

Find the network N with the lower hybridization number such that the input clusters are `explained' by one of the trees displayed by N

subject to constraints on the complexity of N

Phylogenetic network inference

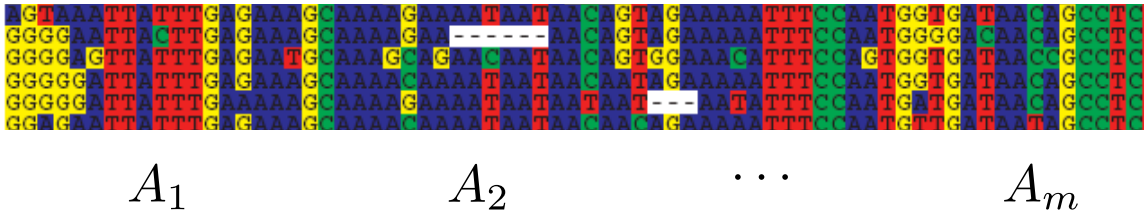
An optimization problem where a candidate network is evaluated *on the basis of how well the trees it displays fit the data:*



Many possible formulations:

Data:

Sequence alignments:
(typically given in blocks)



Goal:

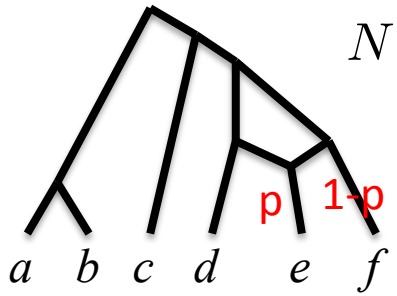
Find N that minimizes $\boldsymbol{F}(N|A_1, A_2, \ldots, A_m) = \sum_{i=1}^{m} \min_{T \in \mathcal{T}(N)} F(T|A_i)$

subject to constraints on the complexity of N. $\boldsymbol{F}()$ is the parsimony score.

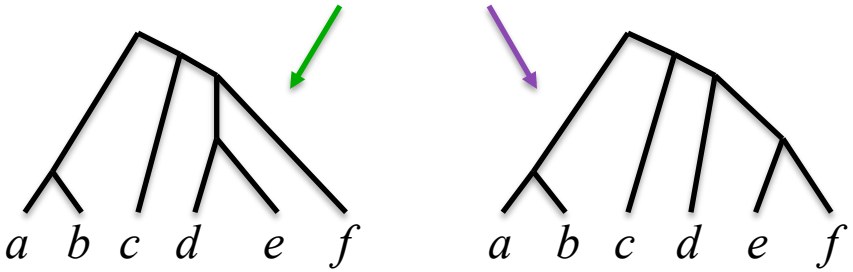
Jin et al. Parsimony Score of Phylogenetic Networks: Hardness Results and a Linear-Time Heuristic. TCCB. 2009.

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:*



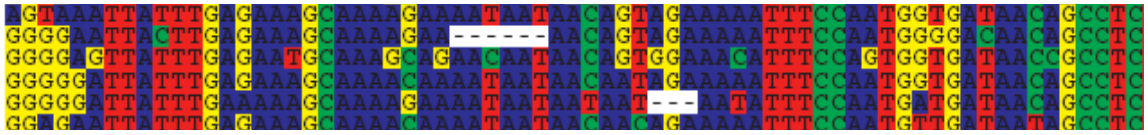
$\mathcal{T}(N)$:



Many possible formulations:

Data:

Sequence alignments:
(typically given in blocks)



A_1 A_2 $\cdots$ A_m

Goal:

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

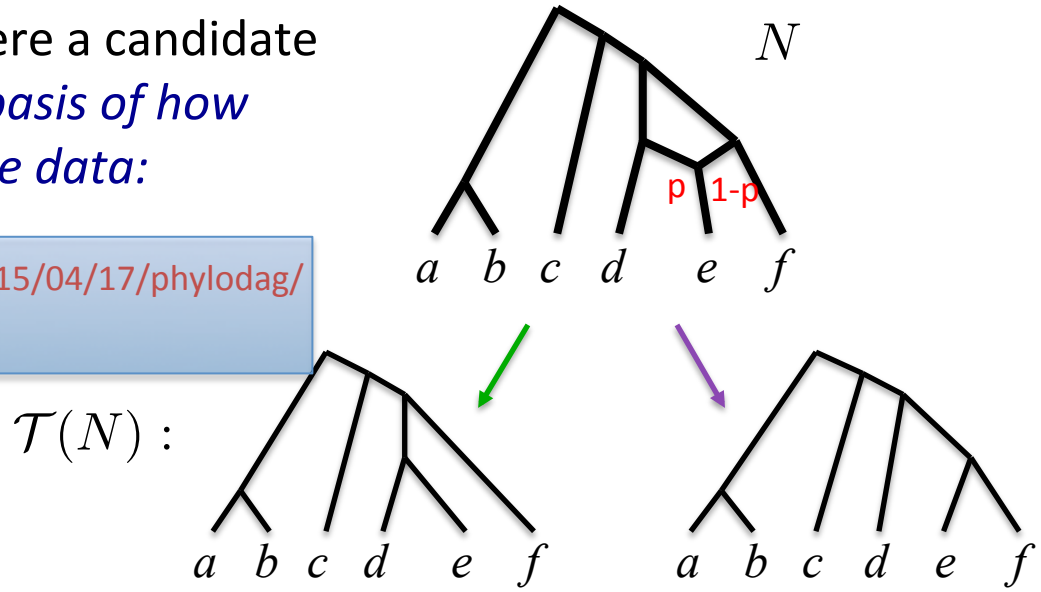
Jin et al.Maximum likelihood of phylogenetic networks. Bioinformatics 2006.

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:*

https://phylomemetic.wordpress.com/2015/04/17/phylodag/

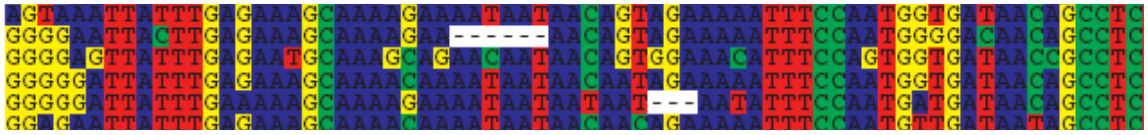
http://old-bioinfo.cs.rice.edu/nepal/



Many possible formulations:

Data:

Sequence alignments:
(typically given in blocks)



A_1 A_2 $\cdots$ A_m

Goal:

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

Jin et al.Maximum likelihood of phylogenetic networks. Bioinformatics 2006.

Some issues

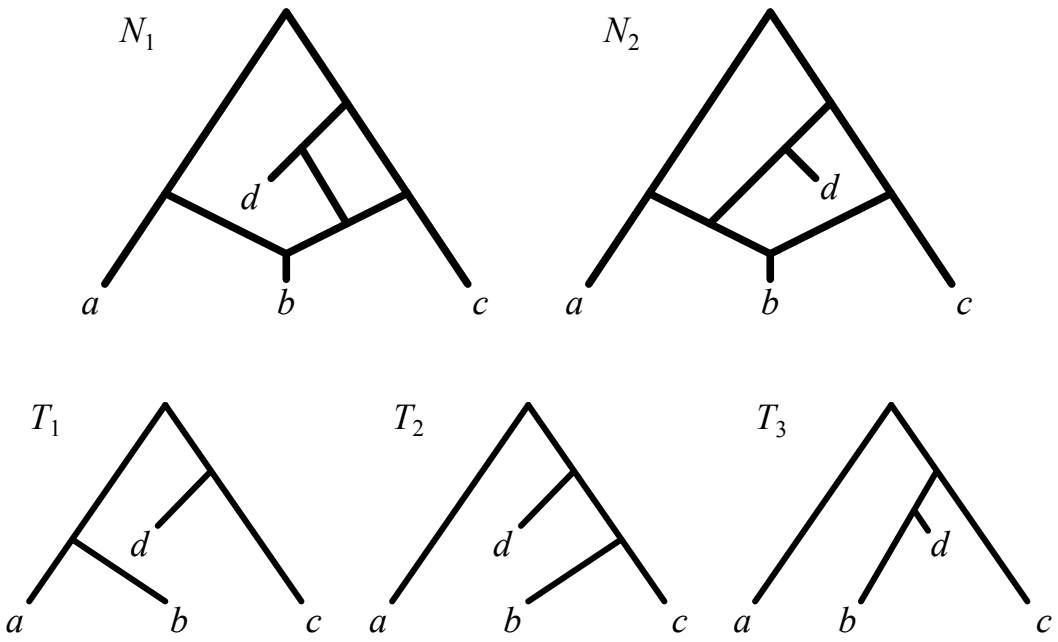
- Searching the space of phylogenetic networks
  The space of networks with k reticulations is infinite.
- Controlling for Model Complexity
  Because any network with k reticulations provides a more complex model than any network with (k-1) reticulations, we must handle the model selection problem (AIC, BIC, K-fold cross-validation, …).
- Identifiability issues

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

- Not accounting for ILS and allopolyploidy

Different networks can display the same trees

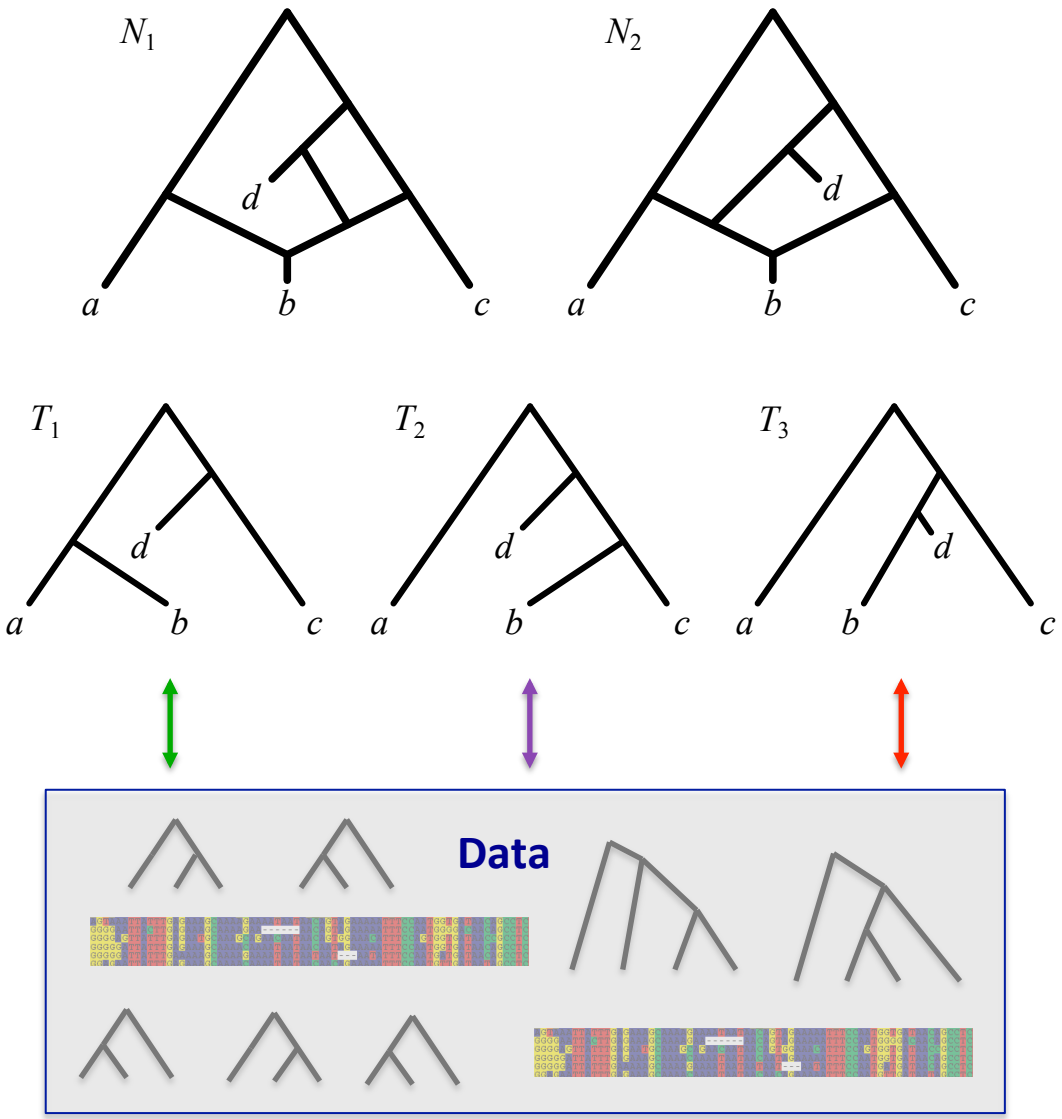
Some networks display exactly the same trees:



Different networks can display the same trees

Some networks display exactly the same trees:

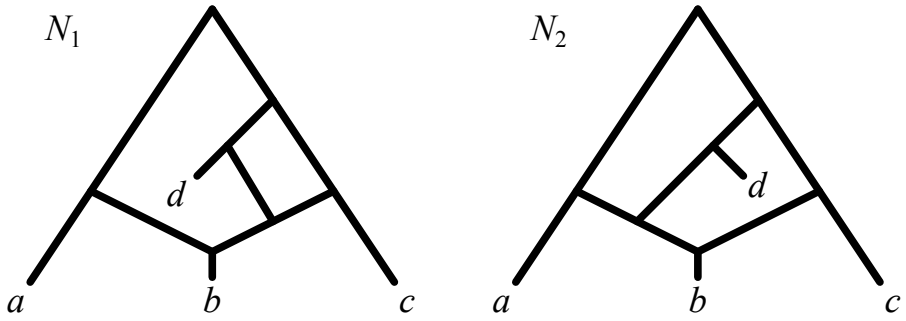
Because N_1 and N_2 display the same trees, they are equally good to any of the inference methods we saw – *no matter the input data*



(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:

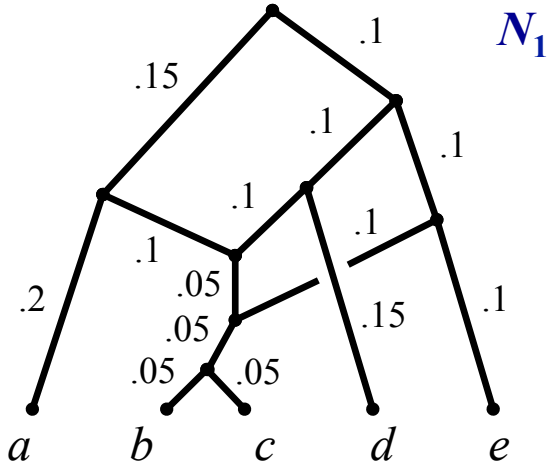


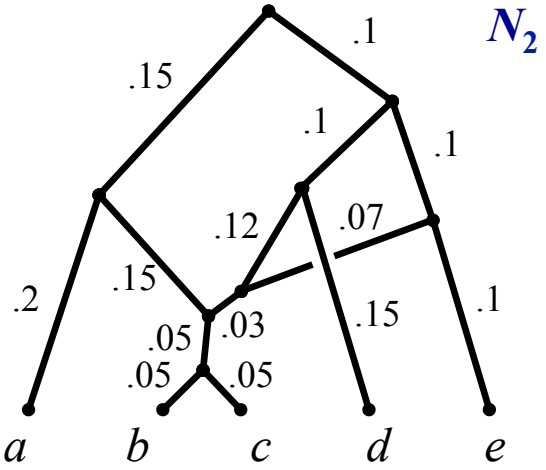
Because N_1 and N_2 display the same trees, they are equally good to any of the inference methods we saw – *no matter the input data*

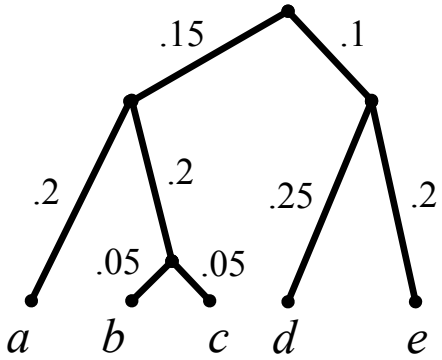
UNIDENTIFIABILITY

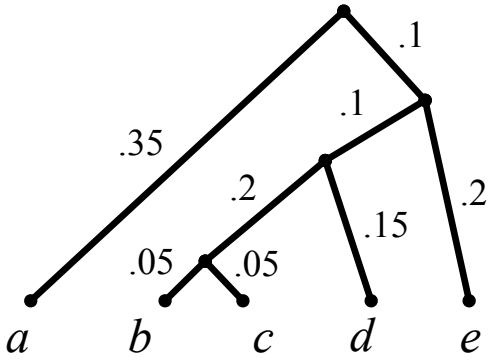
Indistinguishable networks

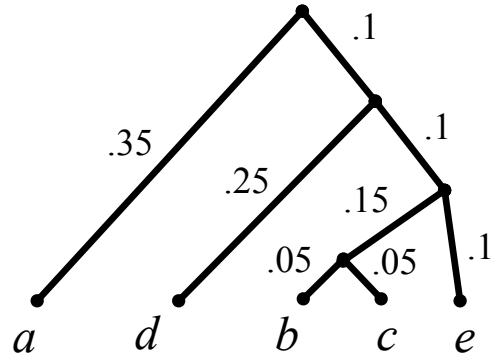
Branch lengths do not eliminate non-identifiability…









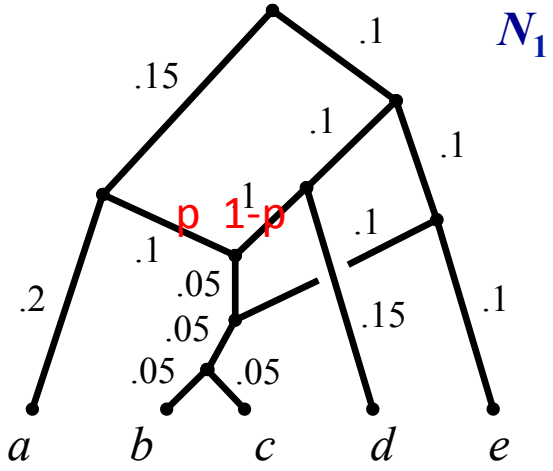


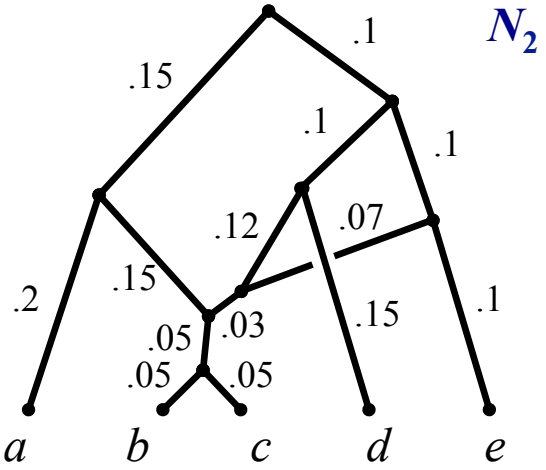
N_1 and N_2 display the same trees (i.e. including branch lengths) and are thus *indistinguishable* even to methods accounting for lengths

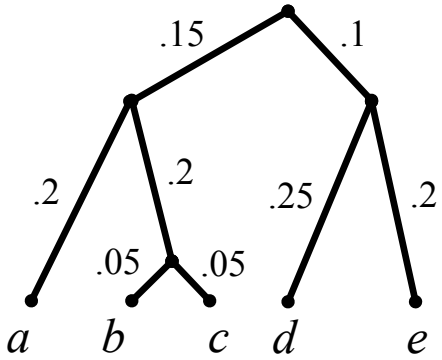
Indistinguishable networks

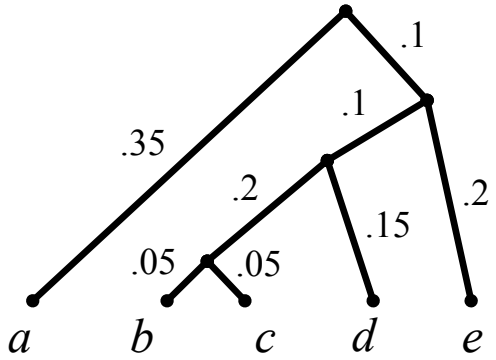
Branch lengths do not eliminate non-identifiability…

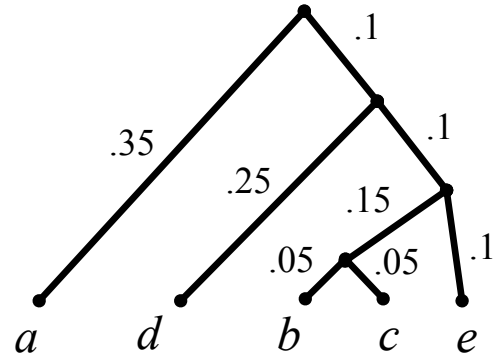
The same hold for inheritance probabilities







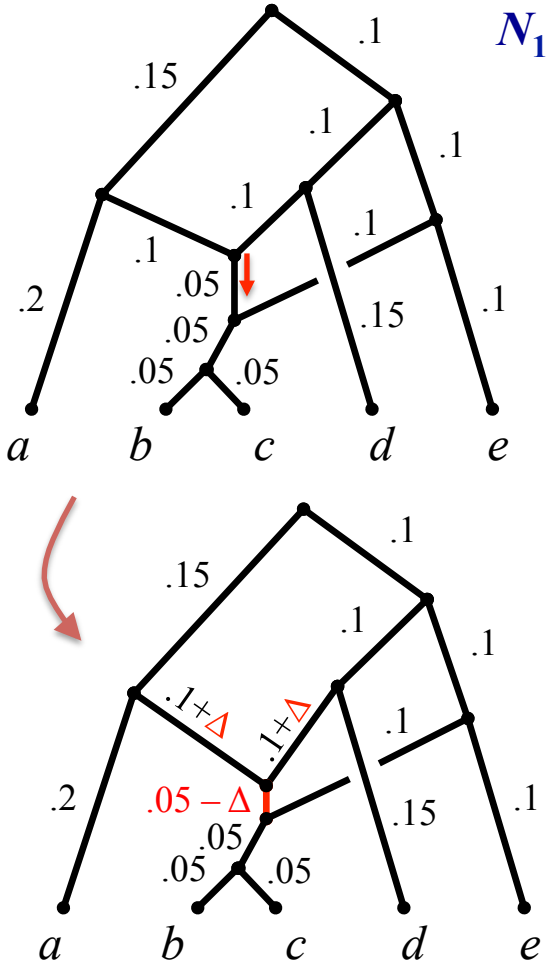




N_1 and N_2 display the same trees (i.e. including branch lengths) and are thus *indistinguishable* even to methods accounting for lengths

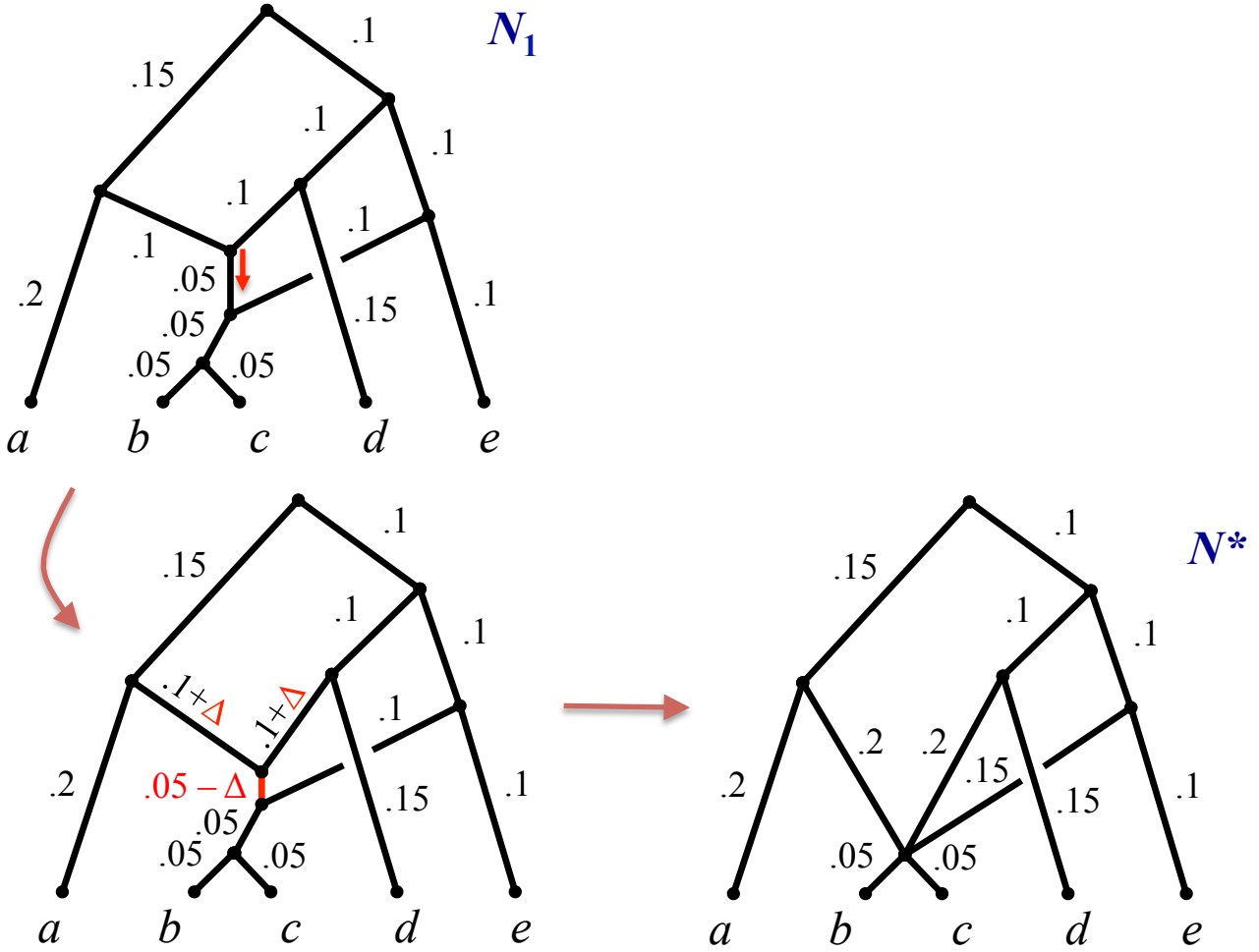
Unzipping a network

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:



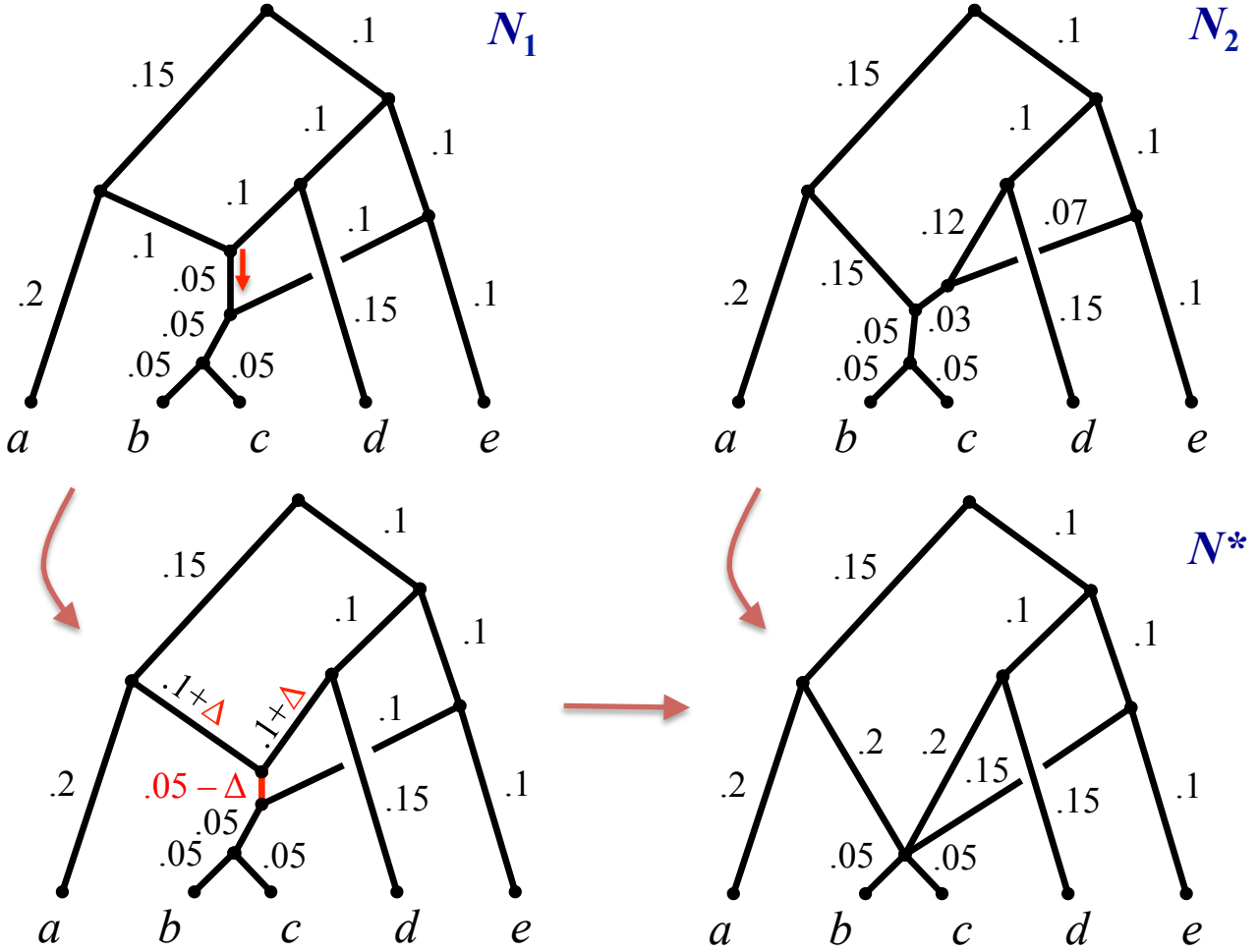
Unzipping a network

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:



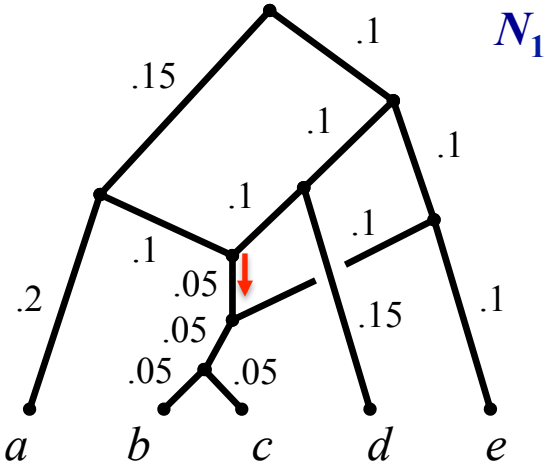
Unzipping a network

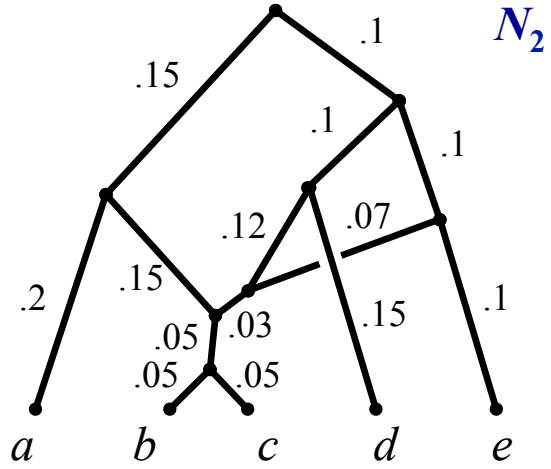
Key observation: we can move reticulations up or down (until they hit a speciation node) and the trees displayed by a network remain the same:



Unzipping a network

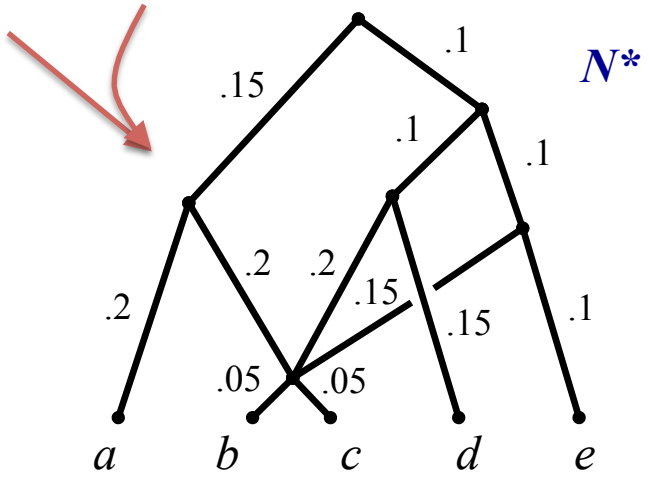
Key observation: we can move reticulations up or down (until they hit a speciation node) and the trees displayed by a network remain the same:





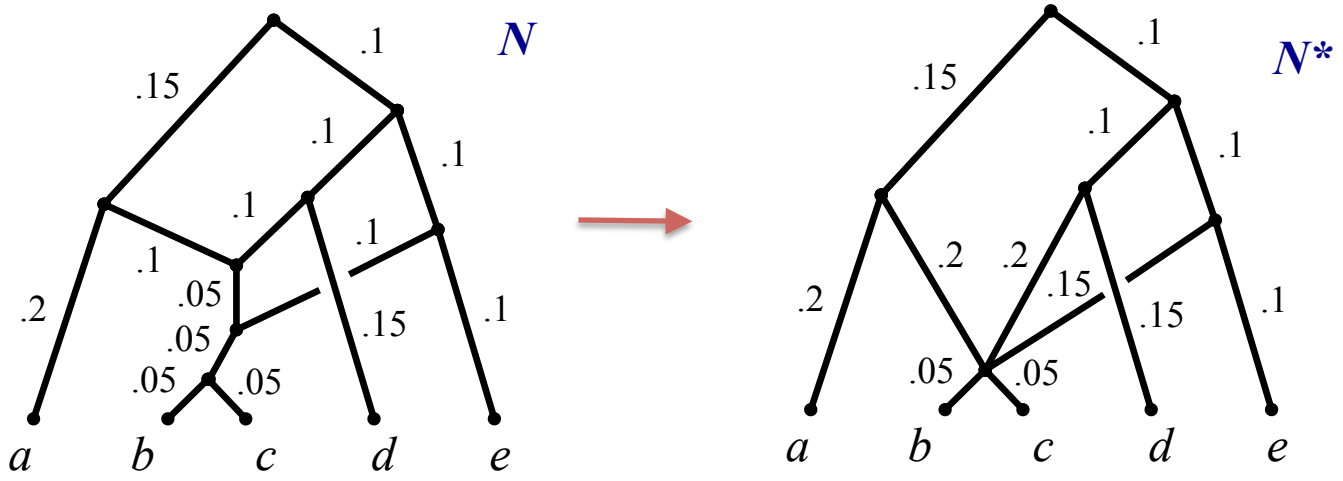
By moving reticulations always down, N_1 and N_2 both end up becoming the same network.

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



What it means for the evolutionary biologist

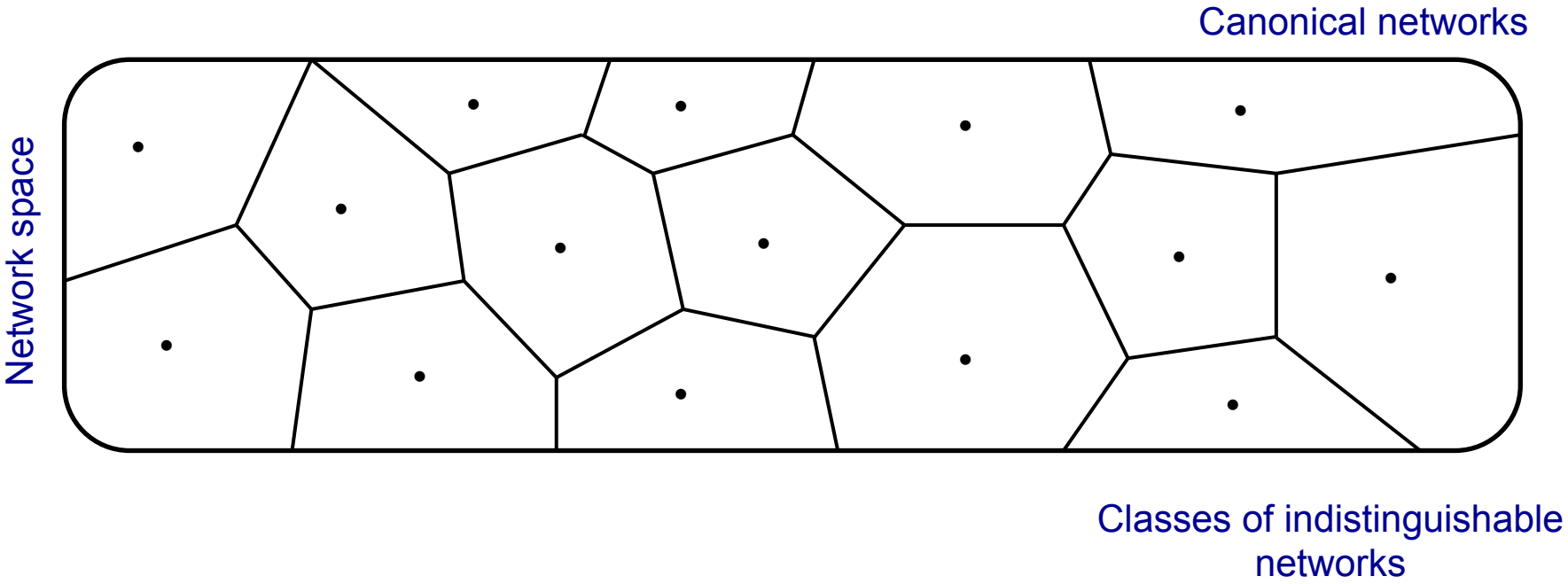
If N is reconstructed by a "classic" inference method, then even assuming perfect and unlimited data, the best you can hope is that the true phylogenetic network is just one of the many that are indistinguishable from N …



The canonical form of N is a unique representative of the networks indistinguishable from N, that excludes their unrecoverable aspects…

Take home message for the computational biologist

If N is reconstructed by a "classic" inference method, then even assuming perfect and unlimited data, the best you can hope is that the true phylogenetic network is just one of the many that are indistinguishable from N …

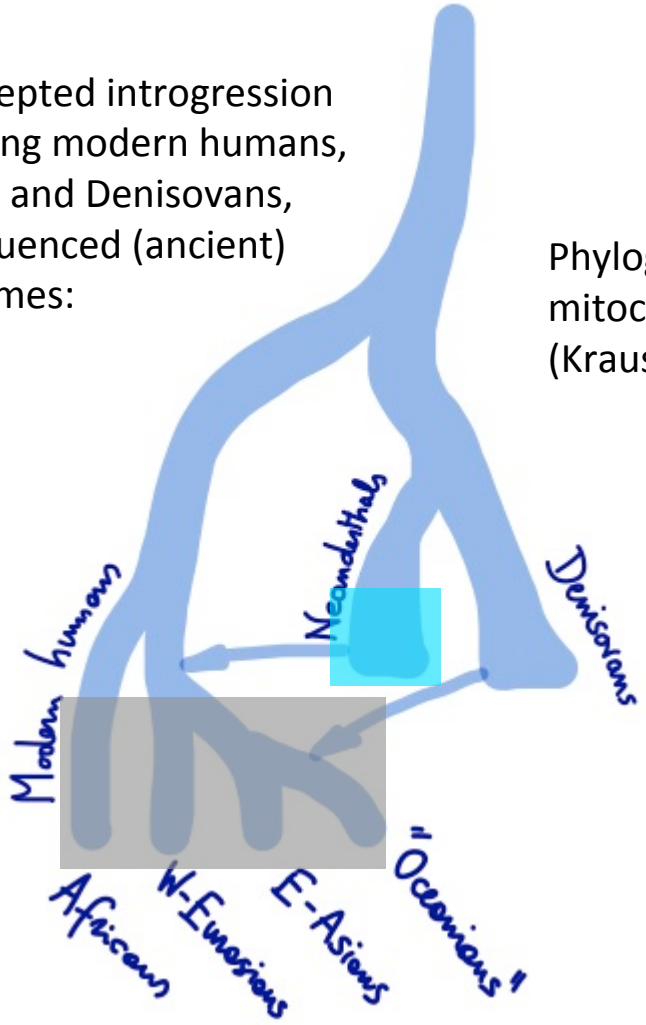


"Classic" network inference methods should only attempt to reconstruct canonical networks .

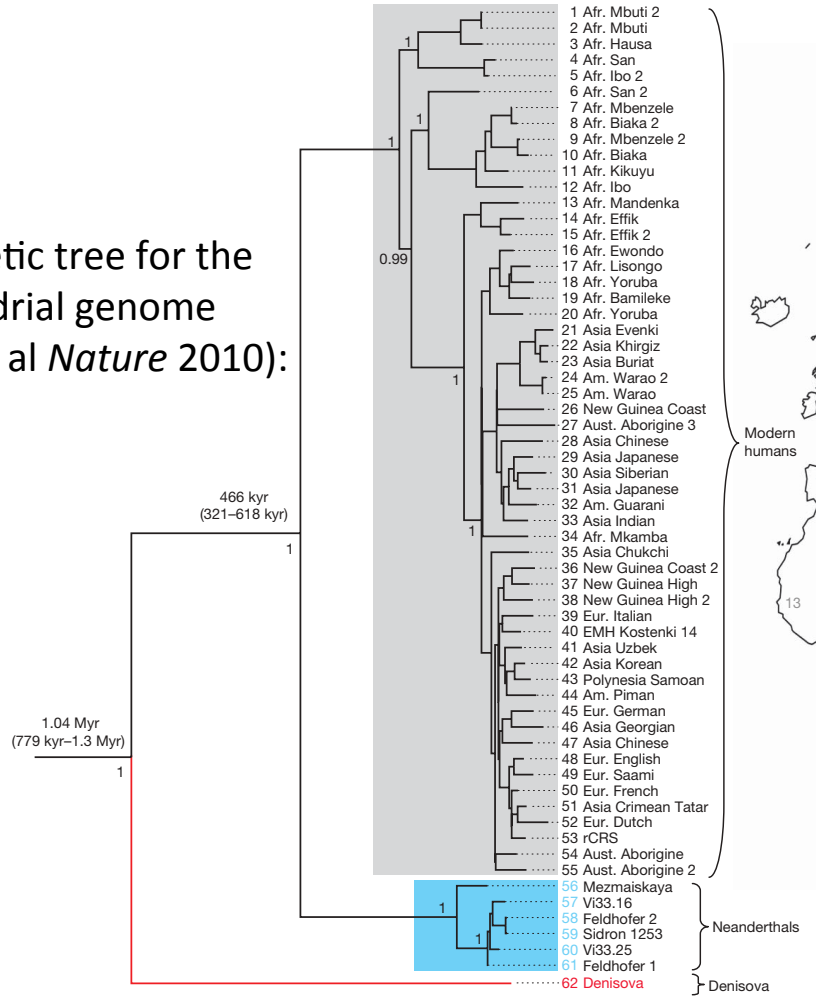
Are gene trees always displayed by the network?

The approaches above assume that **deep coalescence** cannot occur, so gene trees are necessarily displayed trees, but **this is not always the case**

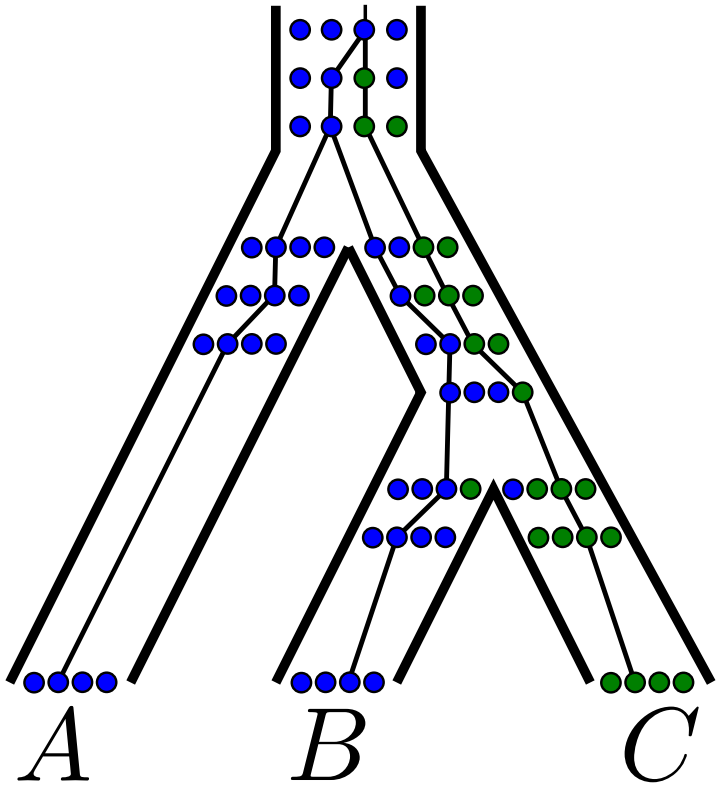
Currently accepted introgression scenario among modern humans, Neanderthals and Denisovans, based on sequenced (ancient) nuclear genomes:



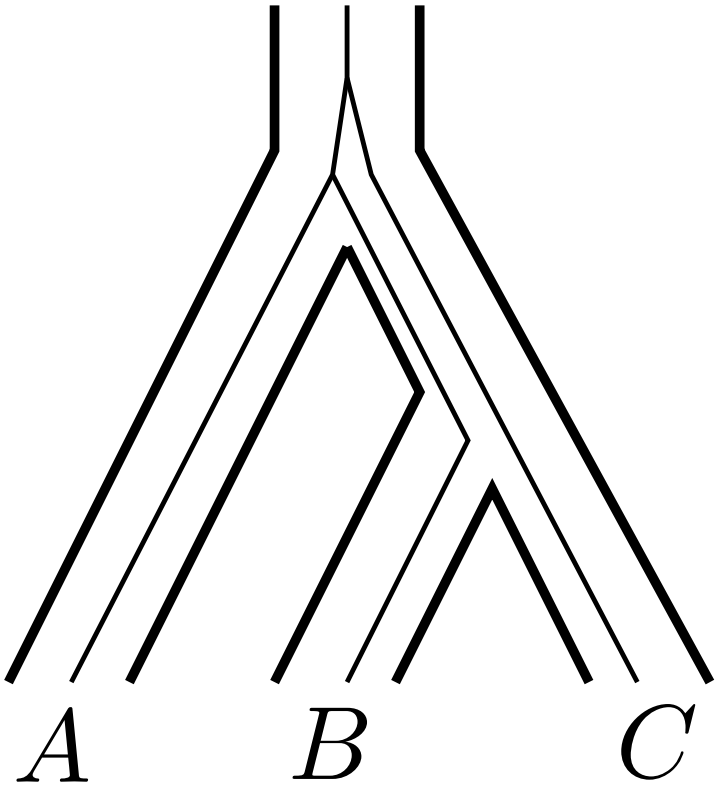
Phylogenetic tree for the mitochondrial genome (Krause et al *Nature* 2010):



An example of incomplete lineage sorting (ILS)

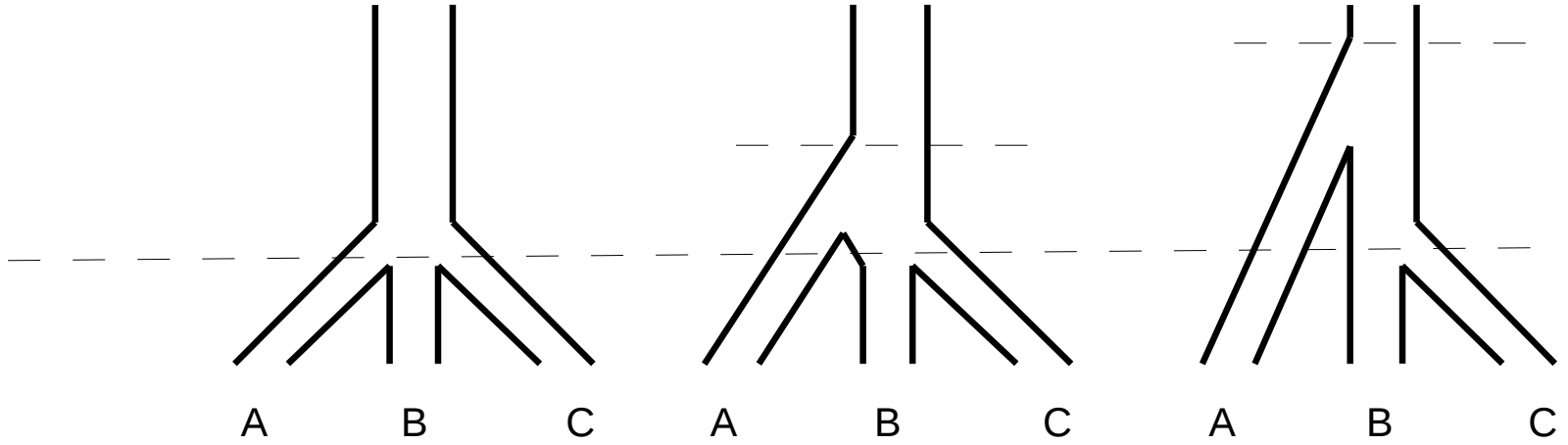


(a) Population view



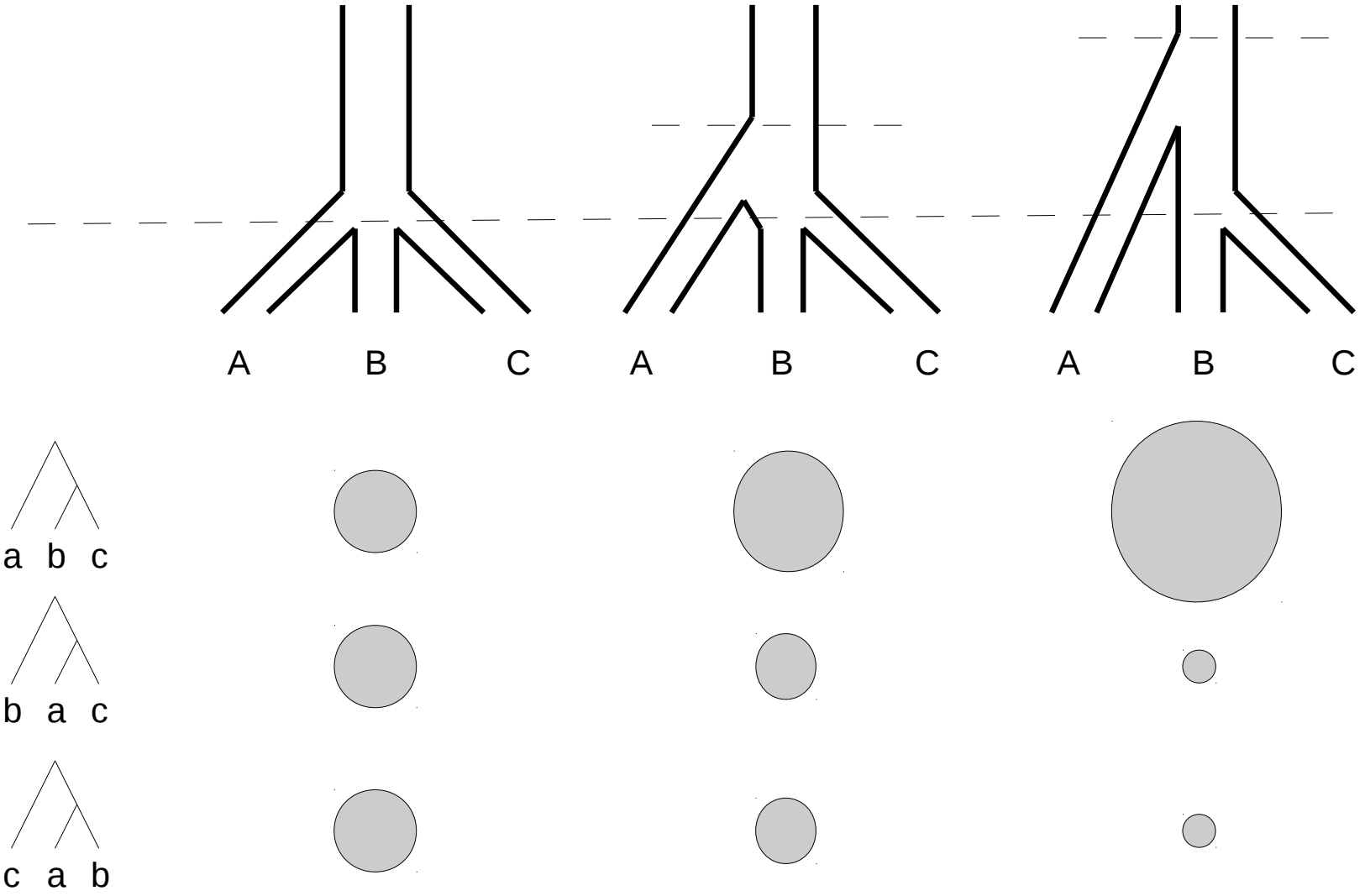
(b) Reconciliation representation

ILS and the probability of gene trees

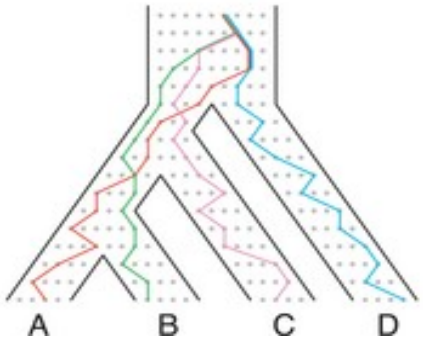


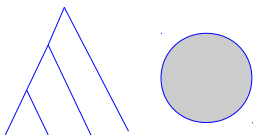
Rannal & Yang 2003 (Genealogies)
Degnan & Salter 2005 (Topologies)

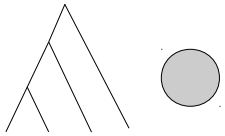
ILS and the probability of gene trees

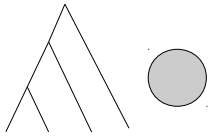


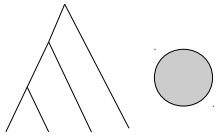
ILS and the probability of gene trees

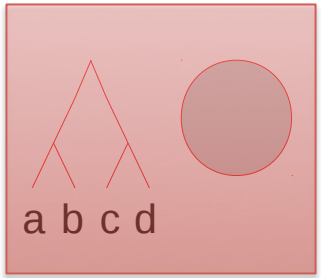


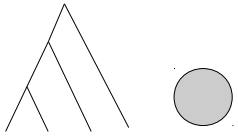

a b c d

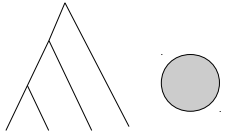

a c d b

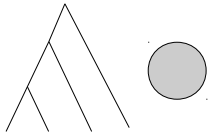

b c a d

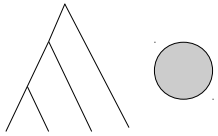

b d c a

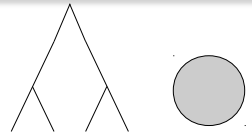

a b c d

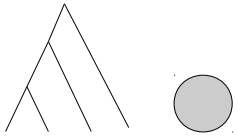

a b d c

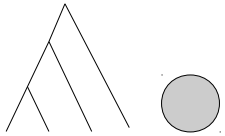

a d b c

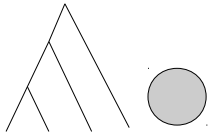

b c d a

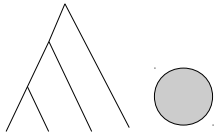

c d a b

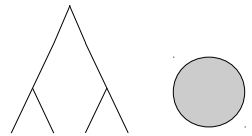

a c b d


a c b d


a d c b


b d a c


c d b a


a d b c

Estimating the species tree under the coalescent model

Since standard phylogenetic methods can be **inconsistent under ILS** (Degnan & Rosenberg 2006), new methods have been developed to cope for this bias, eg:

- STEM (Kubatko, Carstens & Knowles 2009)
- STAR (Liu, Yu, Pearl & Edwards 2009)
- MP-EST (Liu, Yu & Edwards 2010)
- …

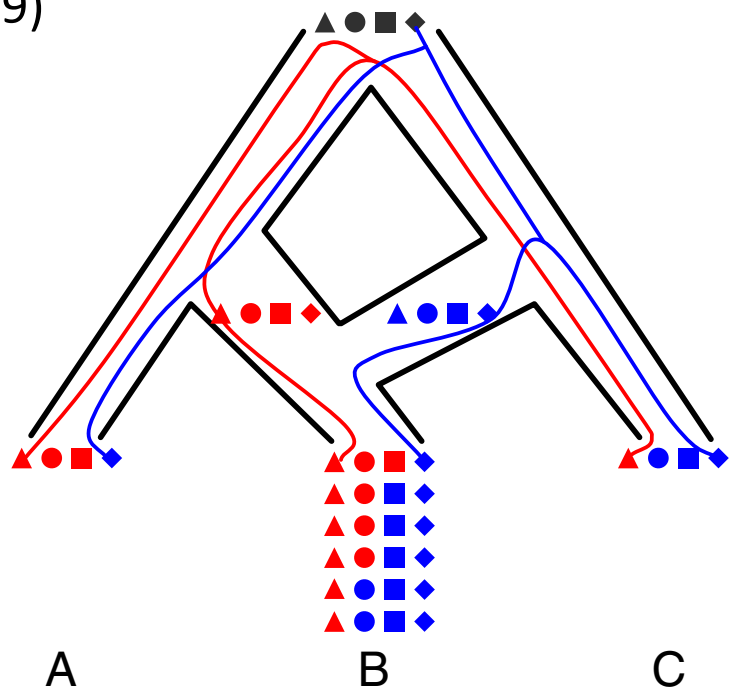
They estimate a phylogenetic species tree given a sample of gene trees (with branch lengths or not) under the coalescent model.

Estimating the species tree under the coalescent model

Since standard phylogenetic methods can be **inconsistent under ILS** (Degnan & Rosenberg 2006), new methods have been developed to cope for this bias, eg:

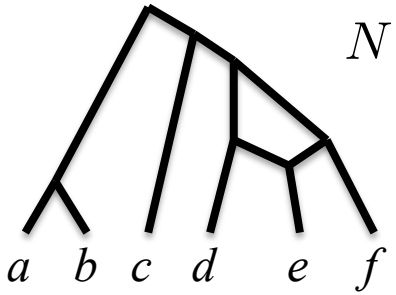
- STEM (Kubatko, Carstens & Knowles 2009)
- STAR (Liu, Yu, Pearl & Edwards 2009)
- MP-EST (Liu, Yu & Edwards 2010)
- …

The same can be done with networks!



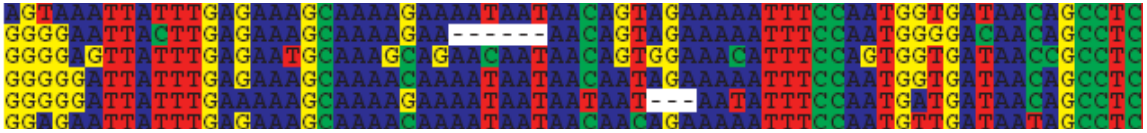
Yu et al. PNAS 2014

Phylogenetic network inference under ILS

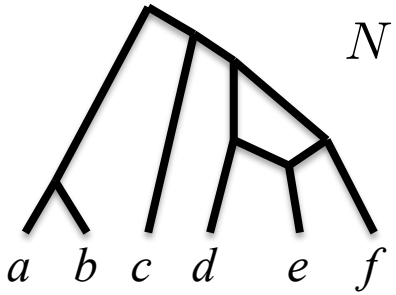


$$\mathbf{Pr}(A_1, A_2, \ldots, A_m|N) = \prod_{i=1}^{m} \mathbf{Pr}(A_i|N) = \prod_{i=1}^{m} \sum_{T \in \mathcal{T}(N)} \mathbf{Pr}(A_i|T)\mathbf{Pr}(T|N).$$

?



Phylogenetic network inference under ILS

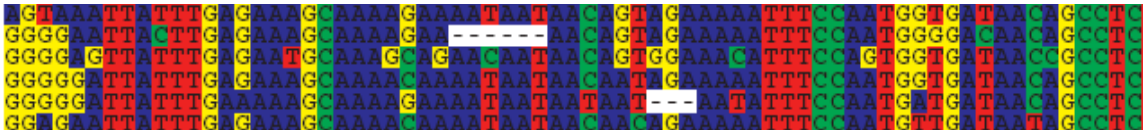


$$\mathbf{Pr}(A_1, A_2, \ldots, A_m|N) = \prod_{i=1}^{m} \mathbf{Pr}(A_i|N) = \prod_{i=1}^{m} \sum_{T \in \mathcal{T}(N)} \mathbf{Pr}(A_i|T)\mathbf{Pr}(T|N).$$

?

$$\mathbf{Pr}(A_1, A_2, \ldots, A_m|N) = \prod_{i=1}^{m} \int_G \mathbf{Pr}(A_i|G)p(G|N).$$

$$\mathbf{Pr}(A_1, A_2, \ldots, A_m|N) = \prod_{i=1}^{m} p(G_i|N).$$



Phylogenetic network inference under ILS

Y Yu & L Nakhleh's group :

- Yu et al. BMC Bioinformatics 2013 (without branch lengths)
- Yu et al. PNAS 2014 (with branch lengths)
- Wen el al. PLOS Genetics 2016 (Bayesian method)

PhyloNet http://bioinfo.cs.rice.edu/phylonet

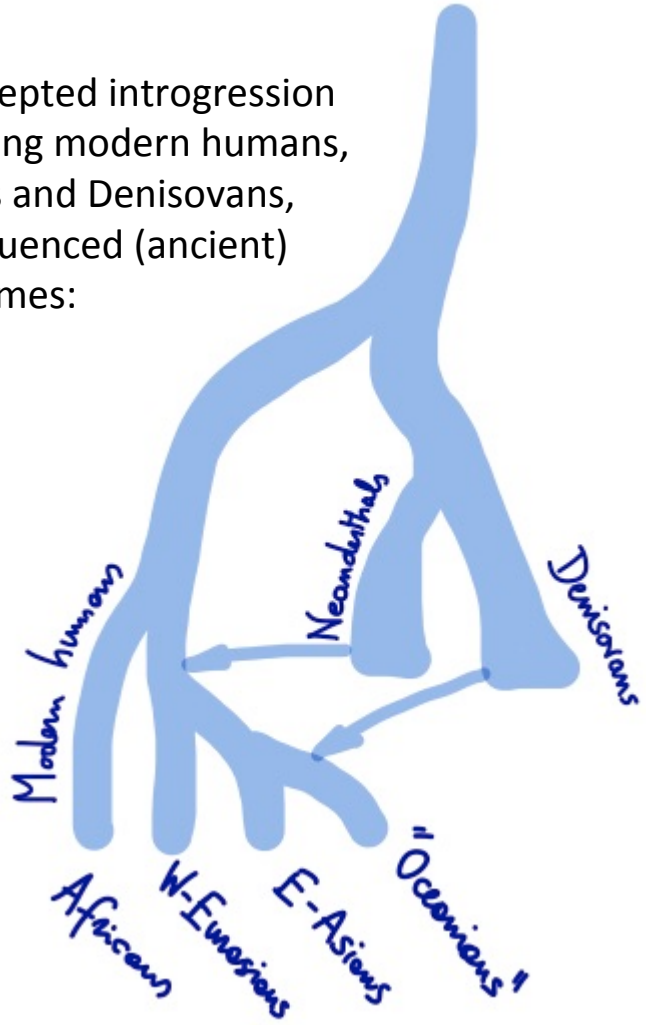
But also other groups contributed significantly, e.g.
L Kubabtko's group and C Ané's group.

$$\mathbf{Pr}(A_1, A_2, \ldots, A_m|N) = \prod_{i=1}^{m} p(G_i|N).$$

Are gene trees always displayed by the network?

The approaches above assume that deep coalescence cannot occur, so gene trees are necessarily displayed trees, but this is not always the case:

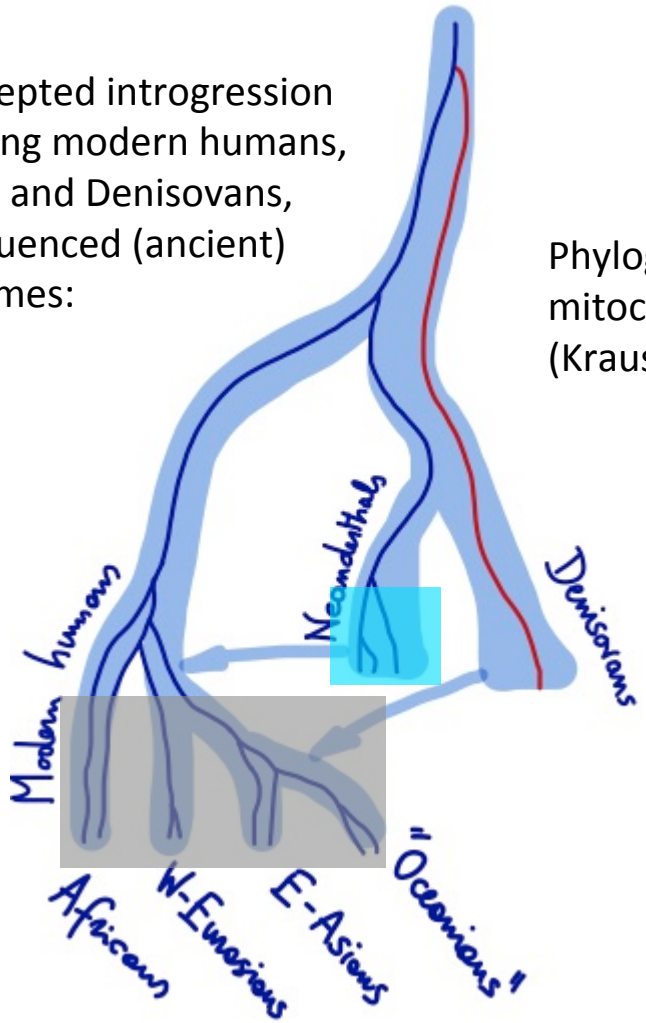
Currently accepted introgression scenario among modern humans, Neanderthals and Denisovans, based on sequenced (ancient) nuclear genomes:



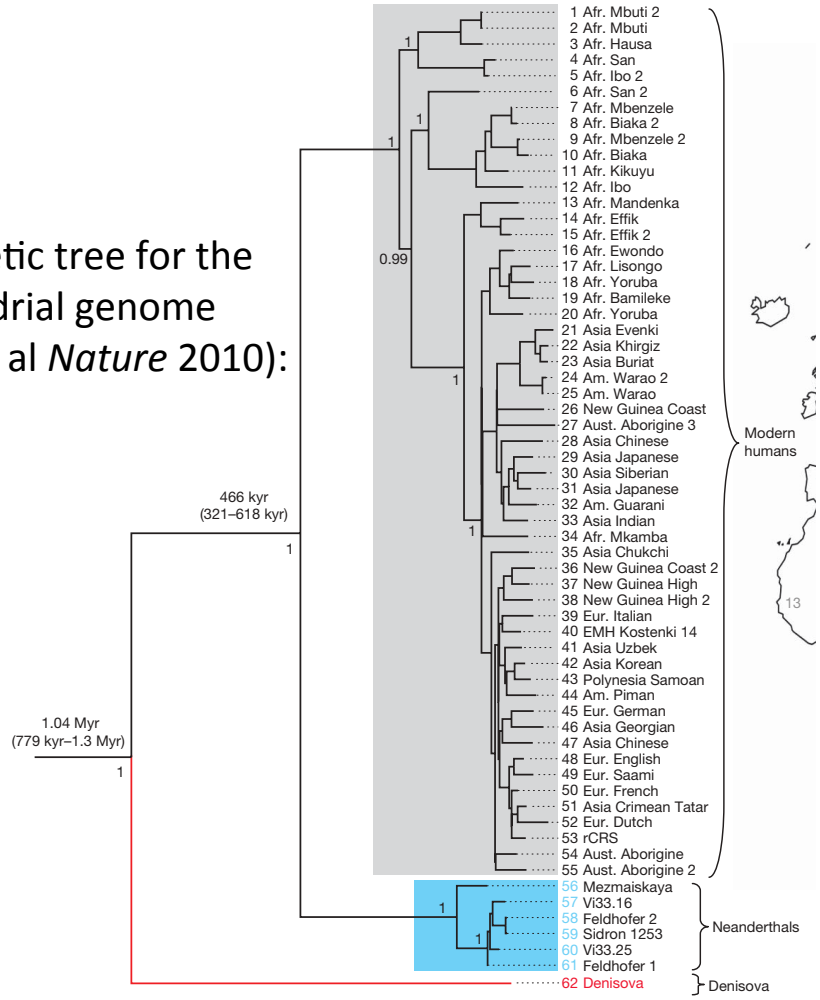
Are gene trees always displayed by the network?

The approaches above assume that deep coalescence cannot occur, so gene trees are necessarily displayed trees, but this is not always the case:

Currently accepted introgression scenario among modern humans, Neanderthals and Denisovans, based on sequenced (ancient) nuclear genomes:

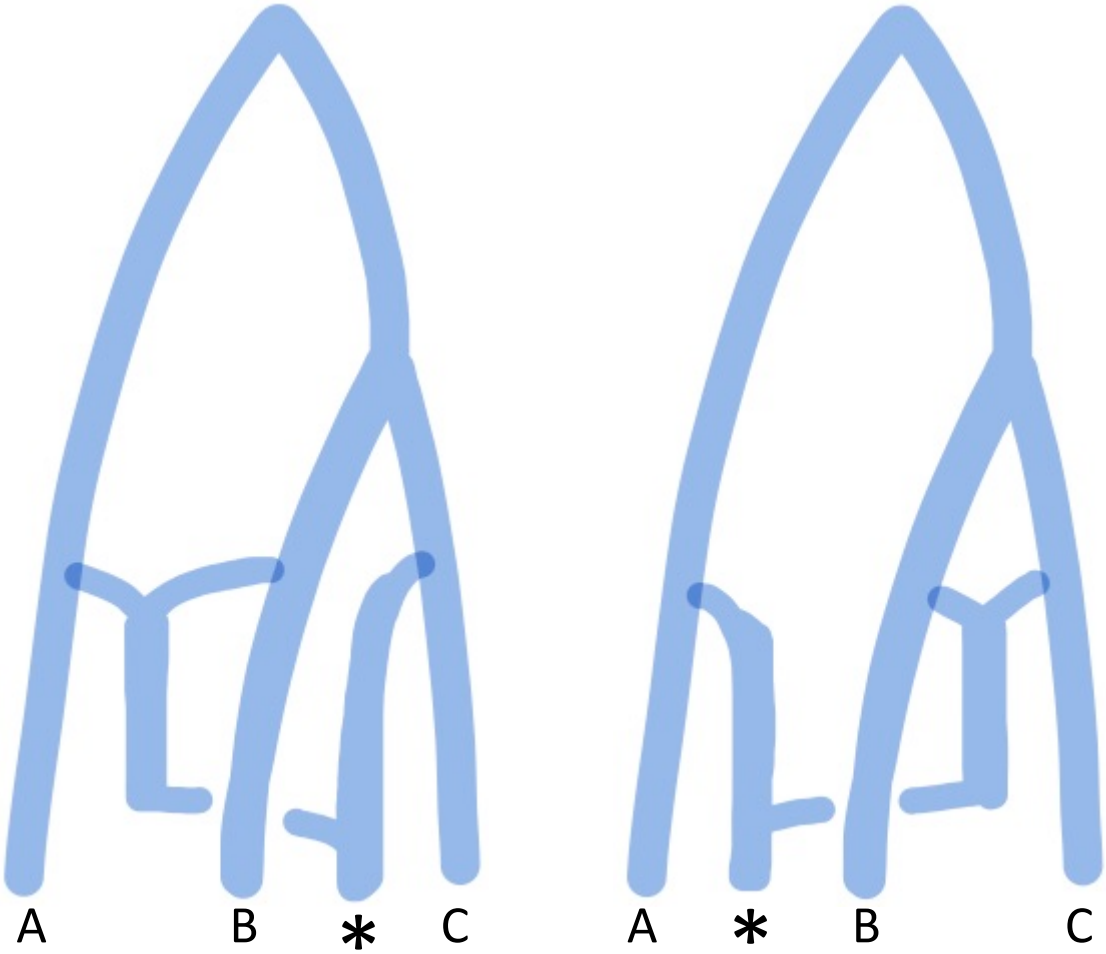


Phylogenetic tree for the mitochondrial genome (Krause et al *Nature* 2010):



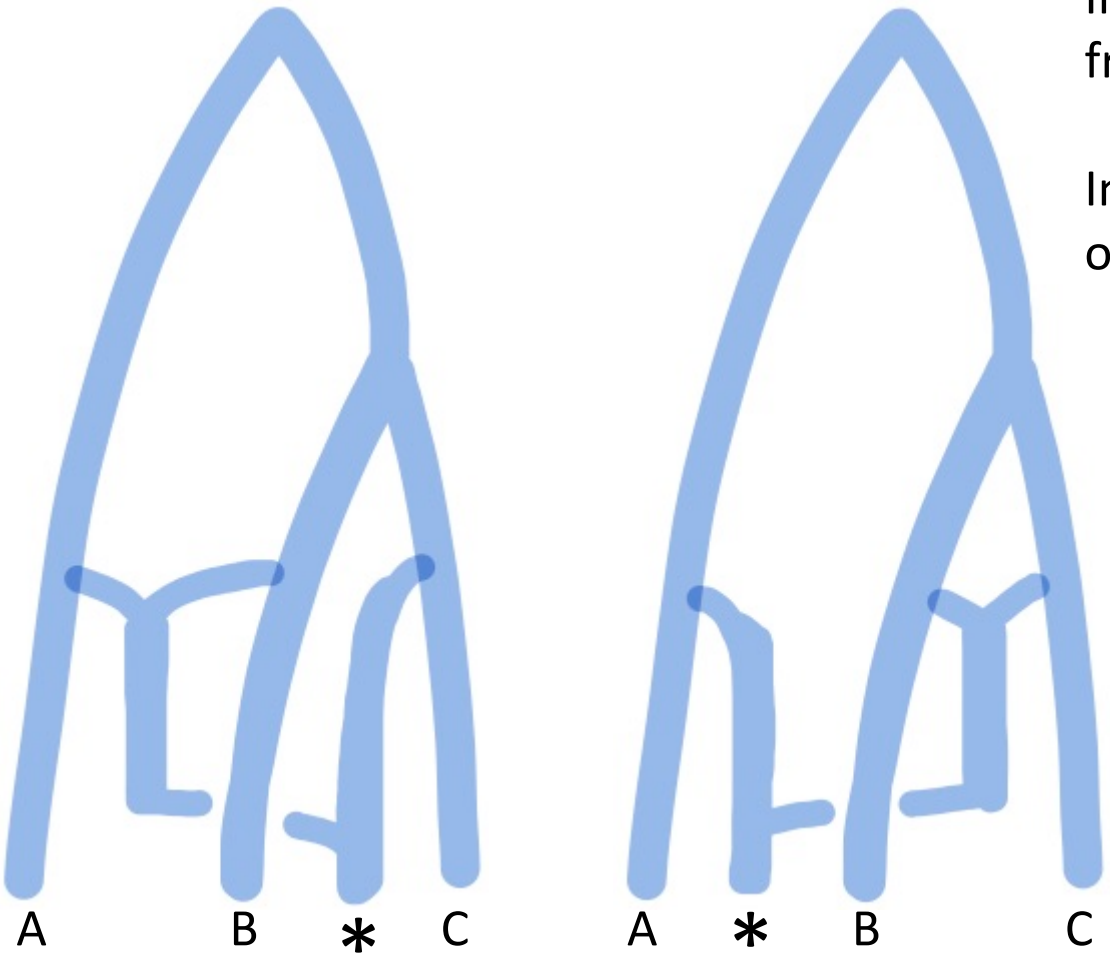
Deep coalescence (ILS) can solve indistinguishability

The following two scenarios would be indistinguishable on the basis of the trees they display:



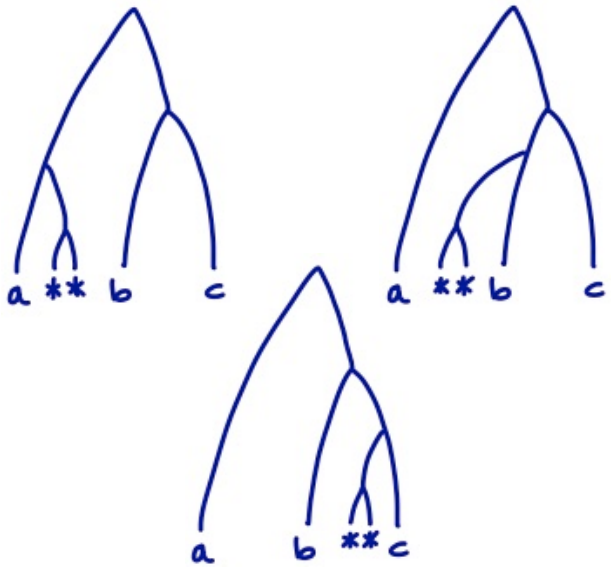
Deep coalescence (ILS) can solve indistinguishability

The following two scenarios would be indistinguishable on the basis of the trees they display:



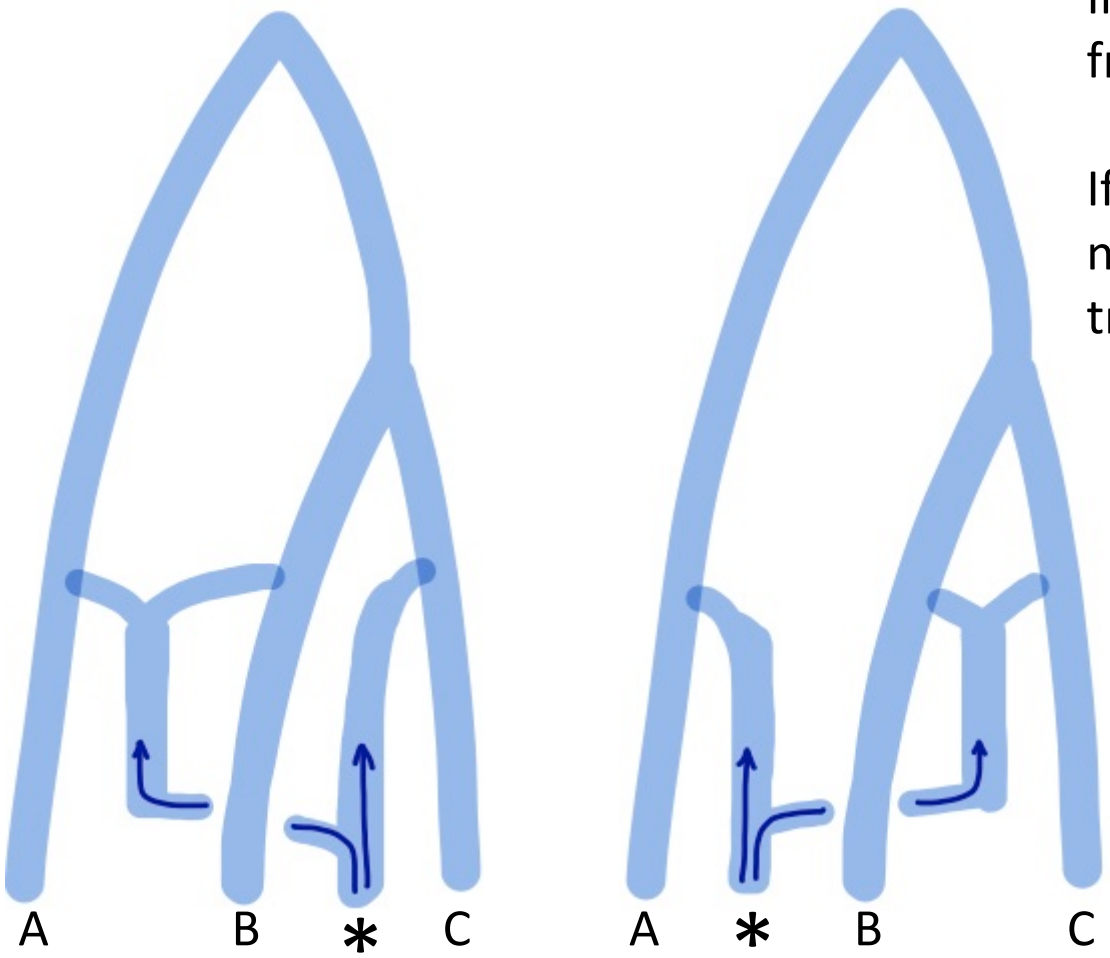
Imagine sampling two alleles from the hybrid species * …

In absence of ILS the only observable gene trees are:



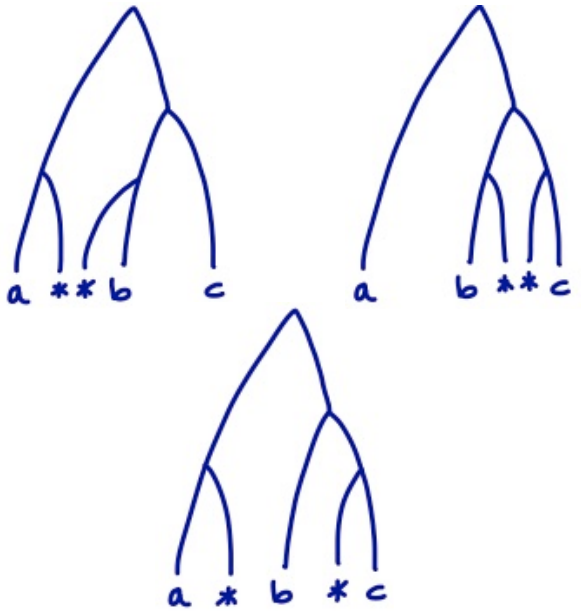
Deep coalescence (ILS) can solve indistinguishability

The following two scenarios would be indistinguishable on the basis of the trees they display:



Imagine sampling two alleles from the hybrid species ∗ …

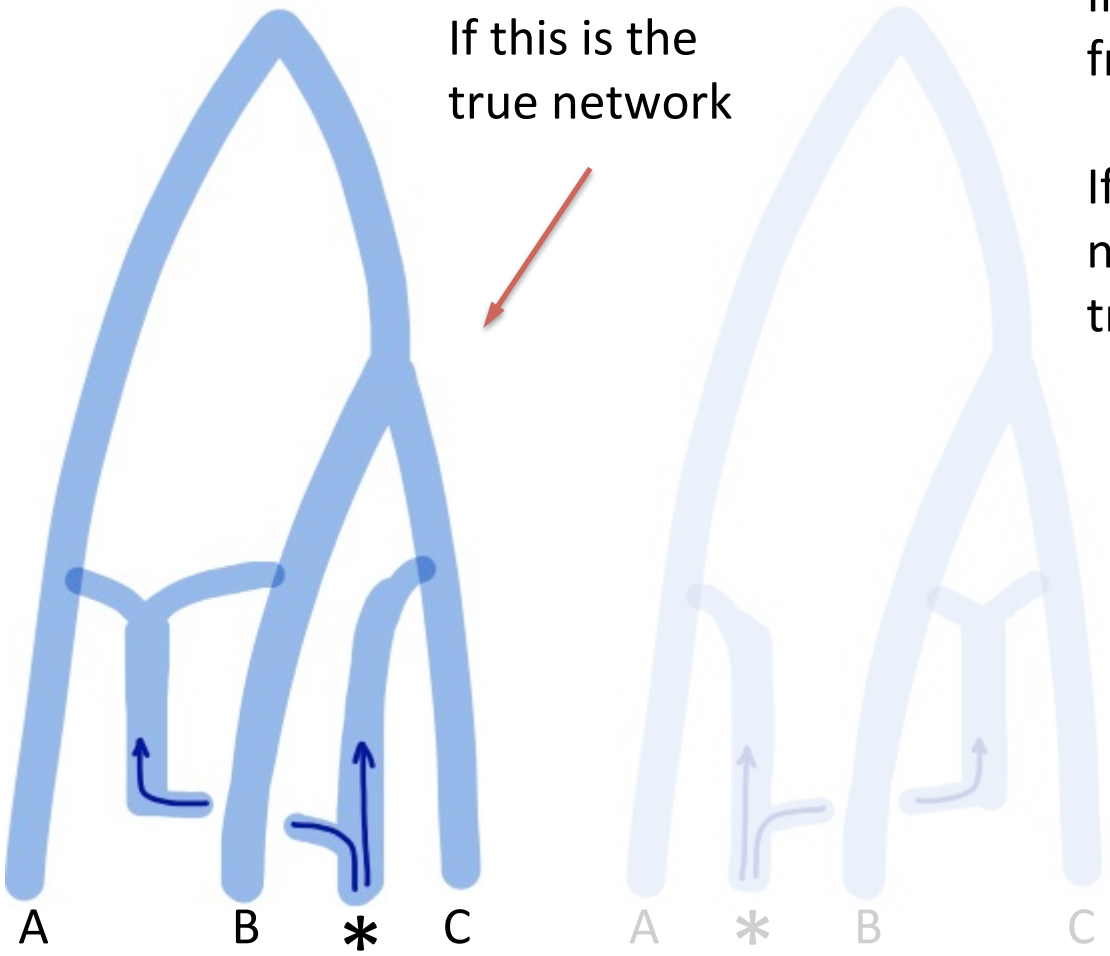
If the two alleles from ∗ do not coalesce, other three gene trees can be observed:



Deep coalescence (ILS) can solve indistinguishability

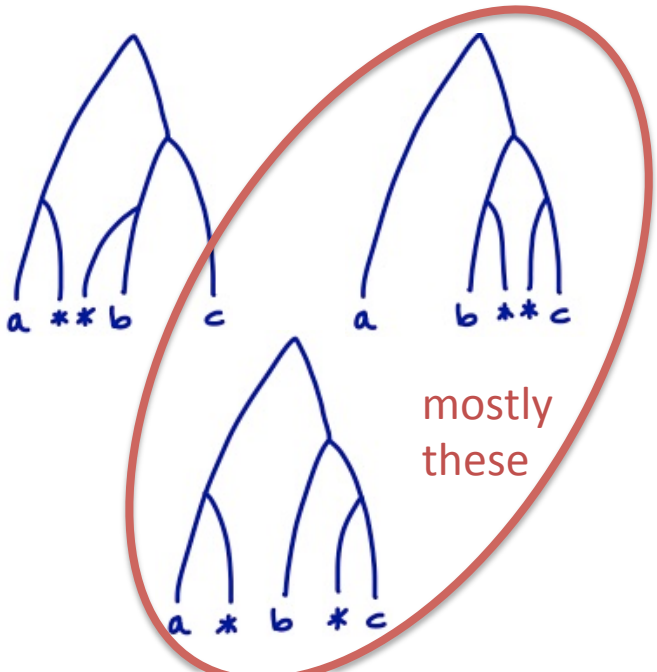
The following two scenarios would be indistinguishable on the basis of the trees they display:

If this is the true network



Imagine sampling two alleles from the hybrid species ∗ …

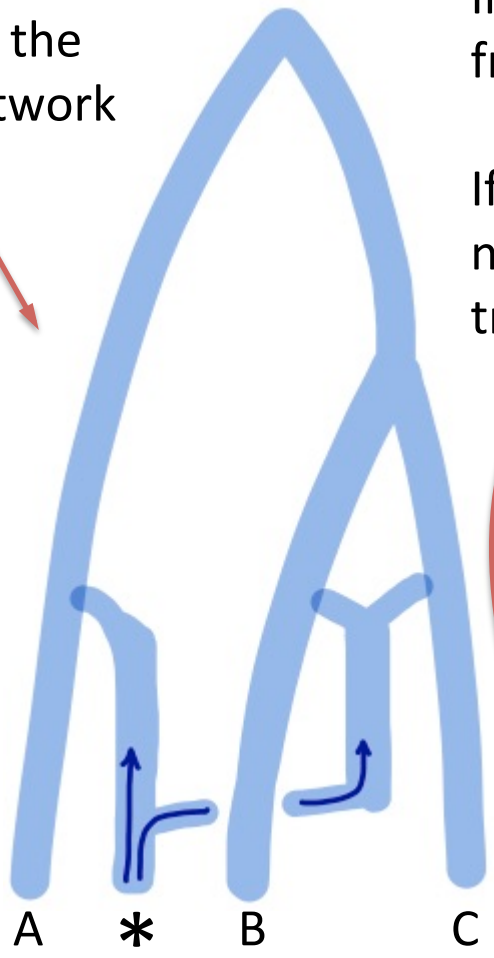
If the two alleles from ∗ do not coalesce, other three gene trees can be observed:



Deep coalescence (ILS) can solve indistinguishability

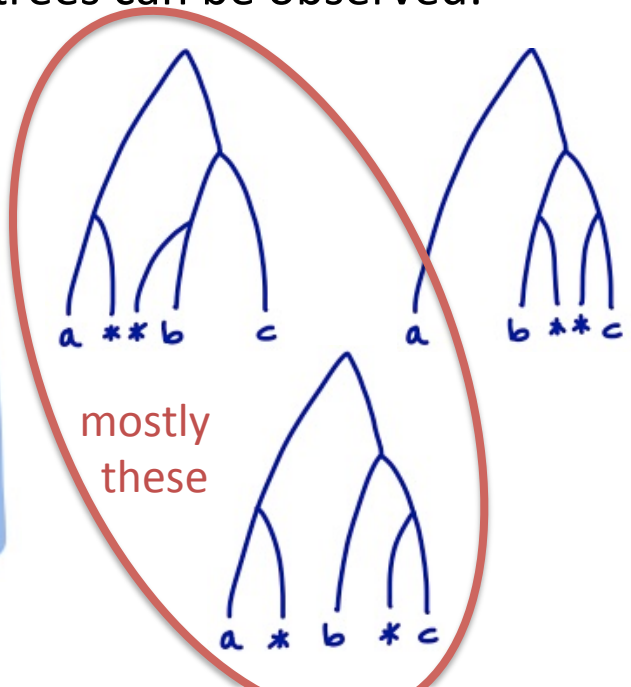
The following two scenarios would be indistinguishable on the basis of the trees they display:

If this is the true network



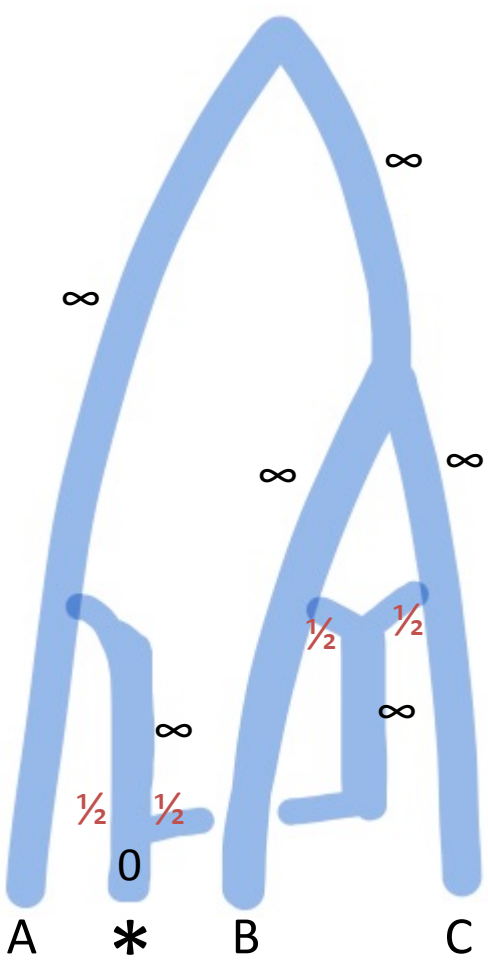
Imagine sampling two alleles from the hybrid species ∗ …

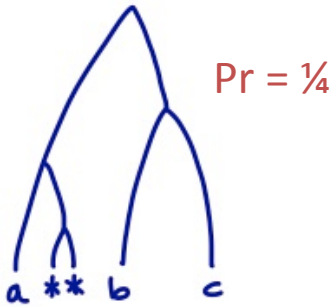
If the two alleles from ∗ do not coalesce, other three gene trees can be observed:



Network Multi-Species Coalescent (NMSC)

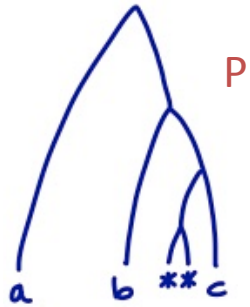
Given a network with branch lengths (in coalescent units) and inheritance probabilities, we can calculate the probability of every possible gene tree:



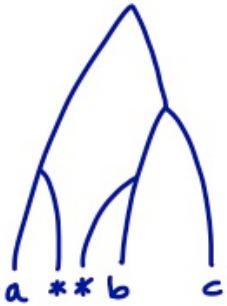


Pr = ¼

Pr = ⅛

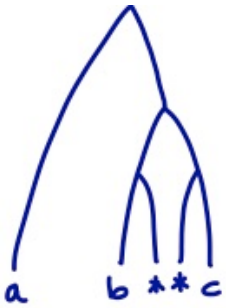


Pr = ⅛



Pr = ¼

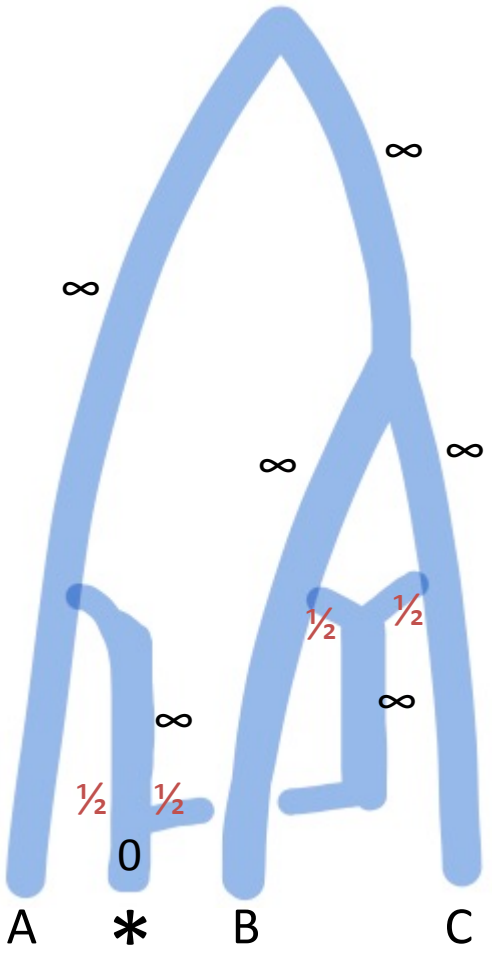
Pr = ¼

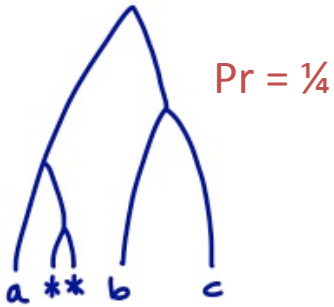


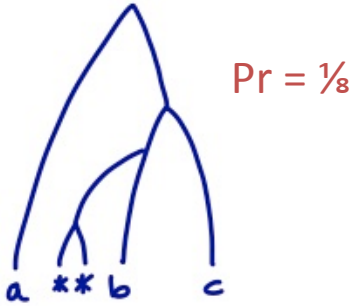
Pr = 0

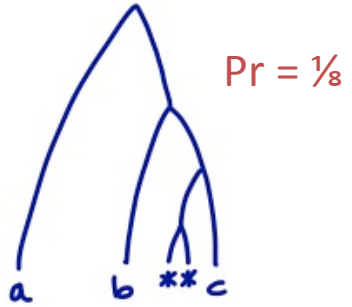
Network Multi-Species Coalescent (NMSC)

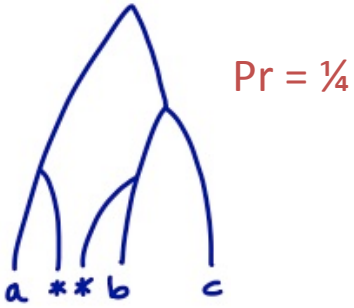
Given a network with branch lengths (in coalescent units) and inheritance probabilities, we can calculate the probability of every possible gene tree:

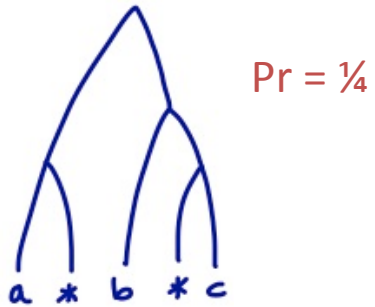








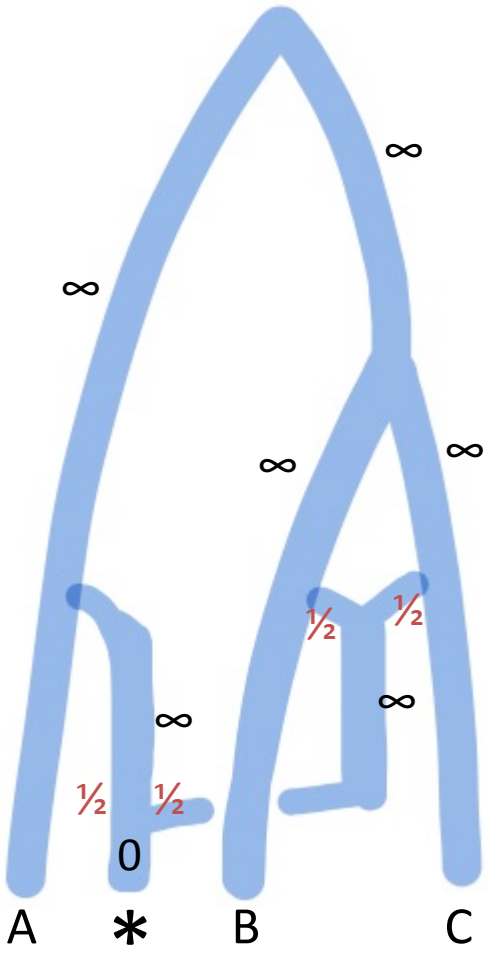


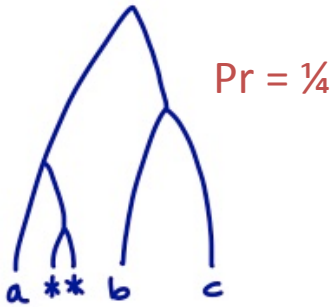


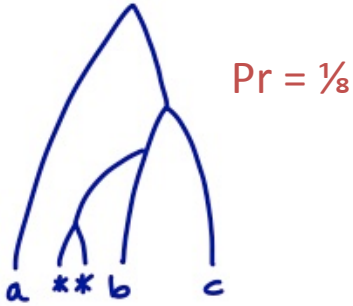
The probability of a gene tree *T* is a weighted average of the probabilities of *T* under a number of "parental trees"… (14 here)

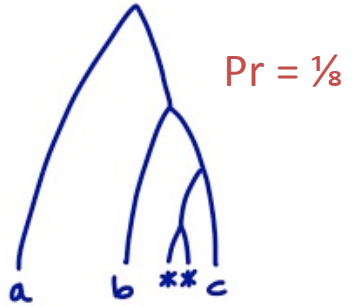
Network Multi-Species Coalescent (NMSC)

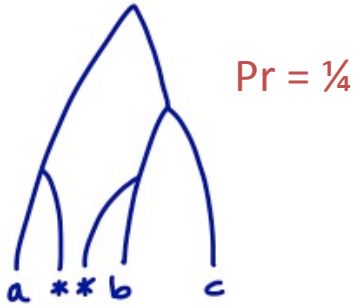
Given a network with branch lengths (in coalescent units) and inheritance probabilities, we can calculate the probability of every possible gene tree:

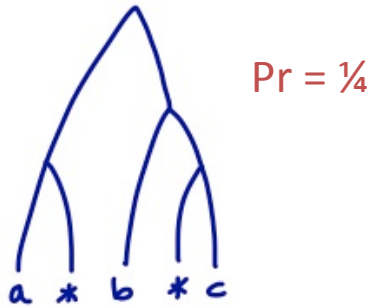












This may solve the identifiability issues for several practical cases.

Note that gene trees can fail to be displayed by the network also because of allopolyploidy.

Some issues

- Searching the space of phylogenetic networks
  The space of networks with k reticulations is infinite.
- Controlling for Model Complexity
  Because any network with k reticulations provides a more complex model than any network with (k-1) reticulations, we must handle the model selection problem (AIC, BIC, K-fold cross-validation, …).
- Identifiability issues
- Not accounting for ILS and allopolyploidy

http://phylnet.univ-mlv.fr/
http://phylonetworks.blogspot.fr

THANKS FOR YOUR ATTENTION