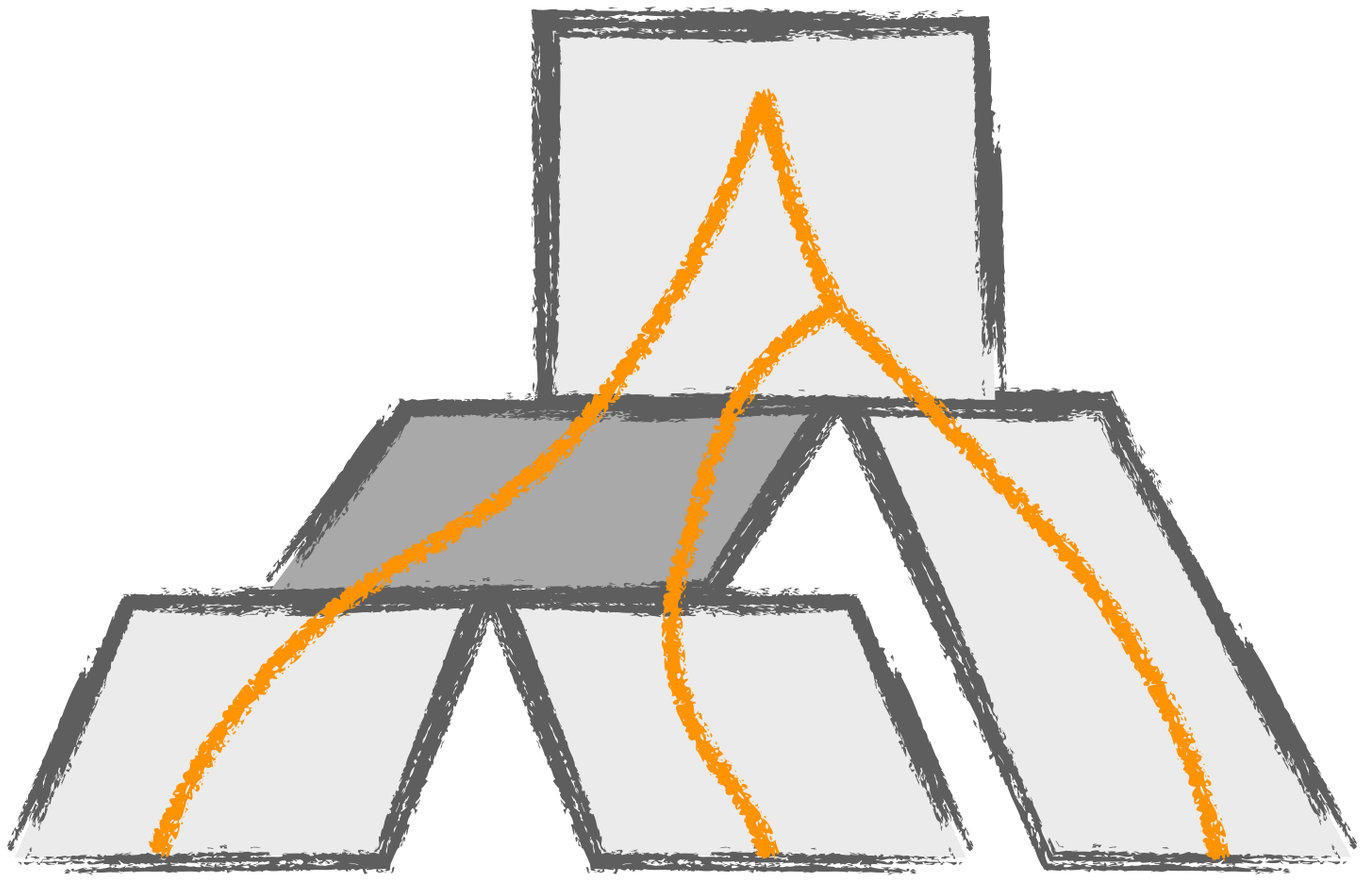
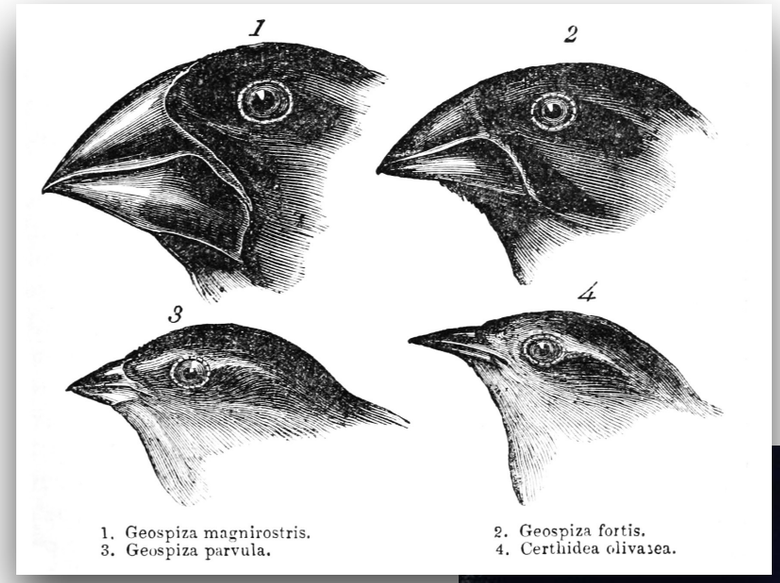
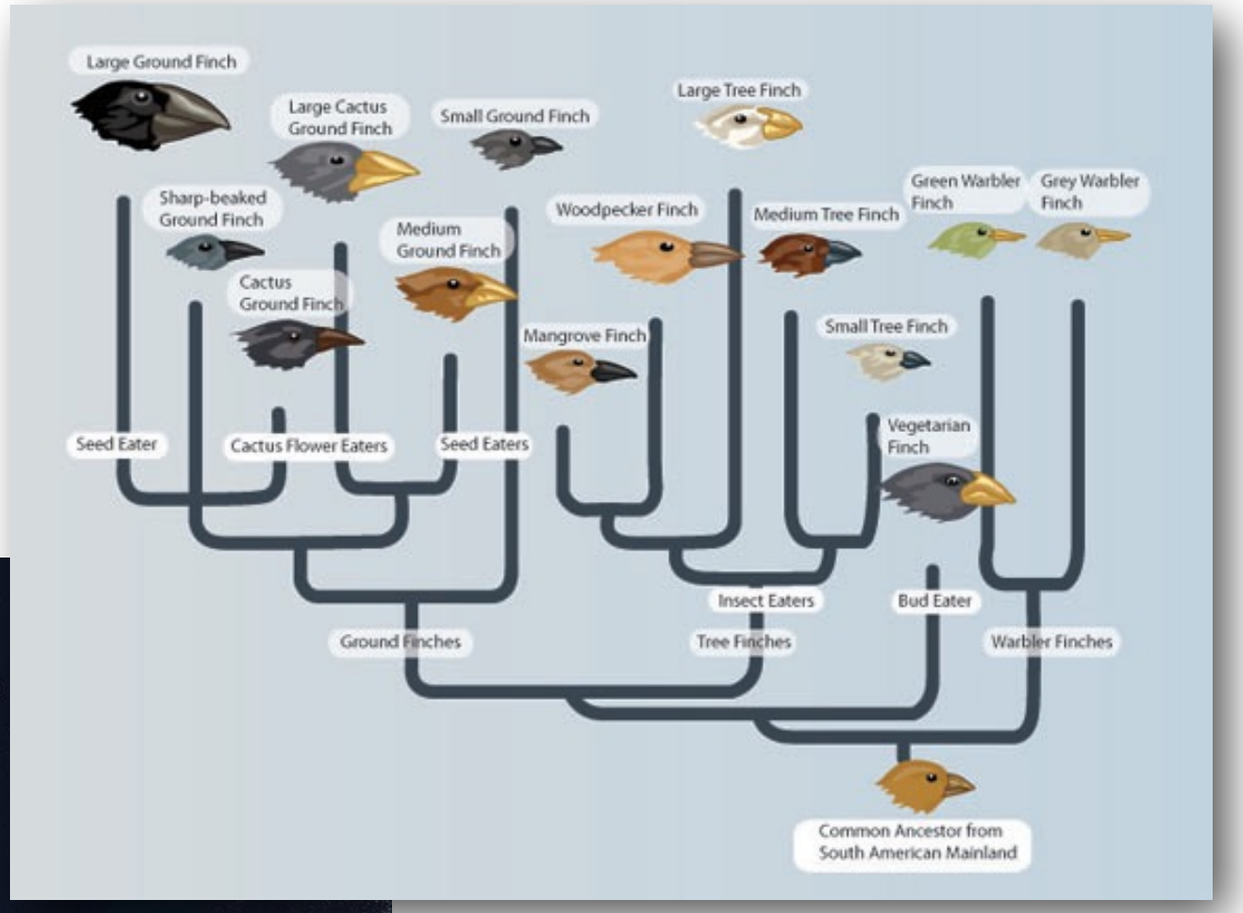


A Biased Walk through Mathematical Phylogenomics

Sébastien Roch
Department of Mathematics
University of Wisconsin-Madison

Darwin's finches





Right figure by Nature Education

Phylogenetic X-trees

Definition

An *X-tree* is a pair $(T;\phi)$ where T is a tree and $\phi : X \to V(T)$ is a labeling such that $\deg(v) \leq 2 \implies v \in \phi(X)$. It is a *phylogenetic X-tree* if ϕ is a bijection into the leaves.

Definition

Two X-trees $(T_1;\phi_1)$ and $(T_2;\phi_2)$ are *isomorphic* if there is a graph isomorphism ψ between T_1 and T_2 such that $\phi_2 = \psi \circ \phi_1$.

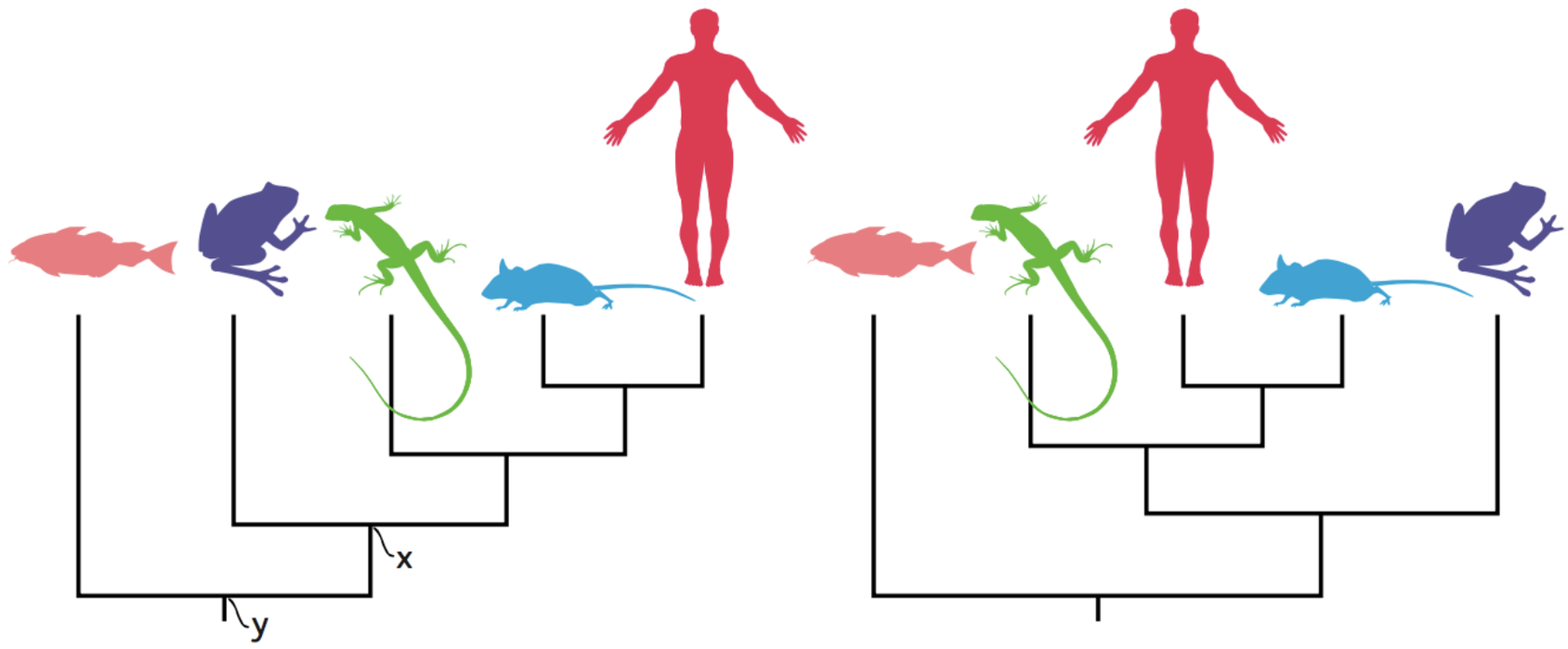
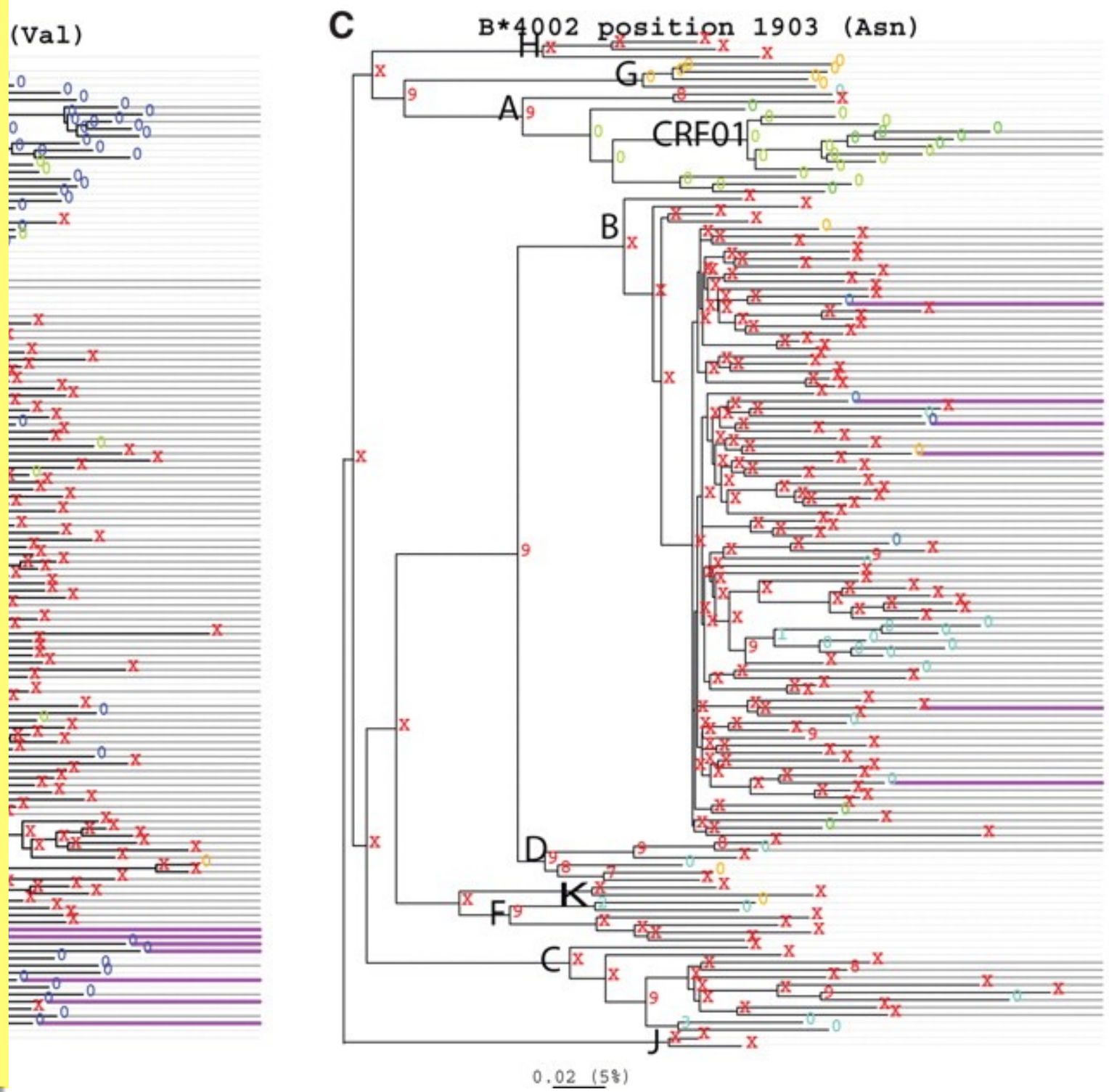


Figure from Baum et al. (2003). Science 310 (5750):979-980.

Why reconstruct phylogenies?



Phylogenetic trees illustrating associations due to subtypes (A) and HLA-driven escape (C).
(From: Tanmoy Bhattacharya et al. Science 2007;315:1583-1586)

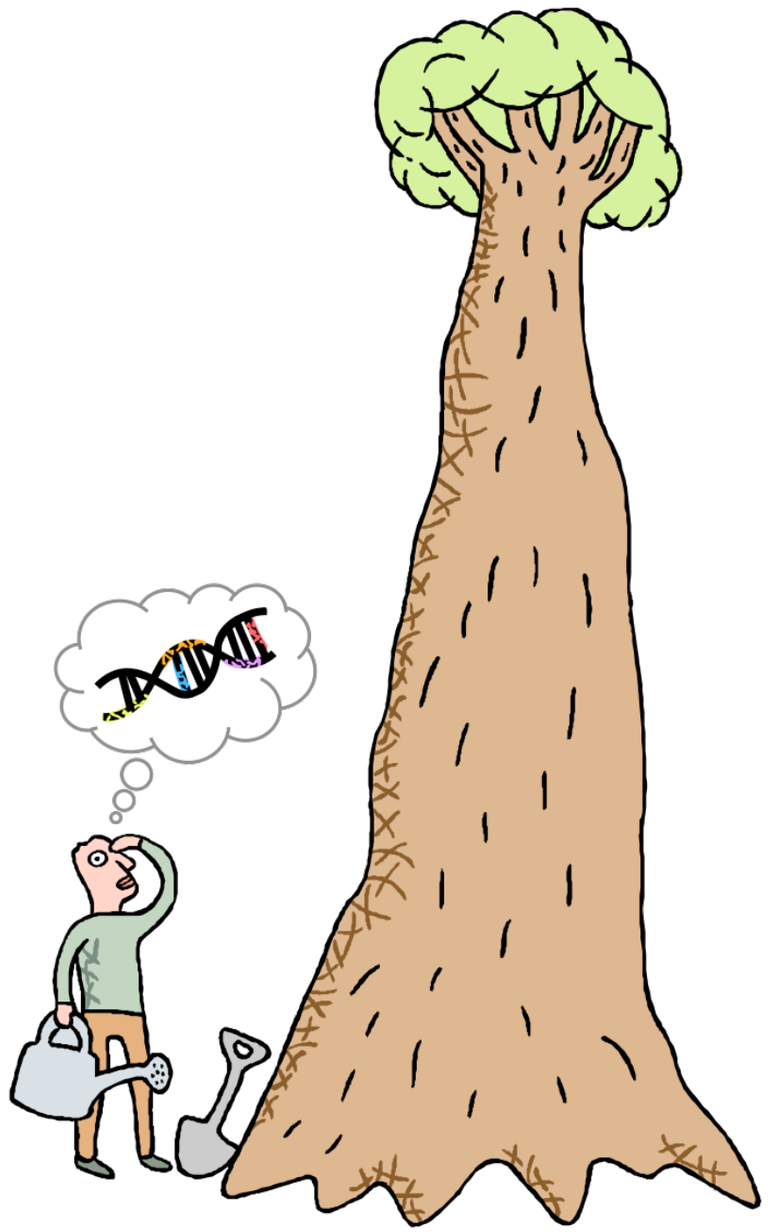
Estimating the Tree of Life: Mathematical challenges in phylogenomics

I. Background:
pre-genomics era

II. More data, more problems:

A. The multispecies coalescent

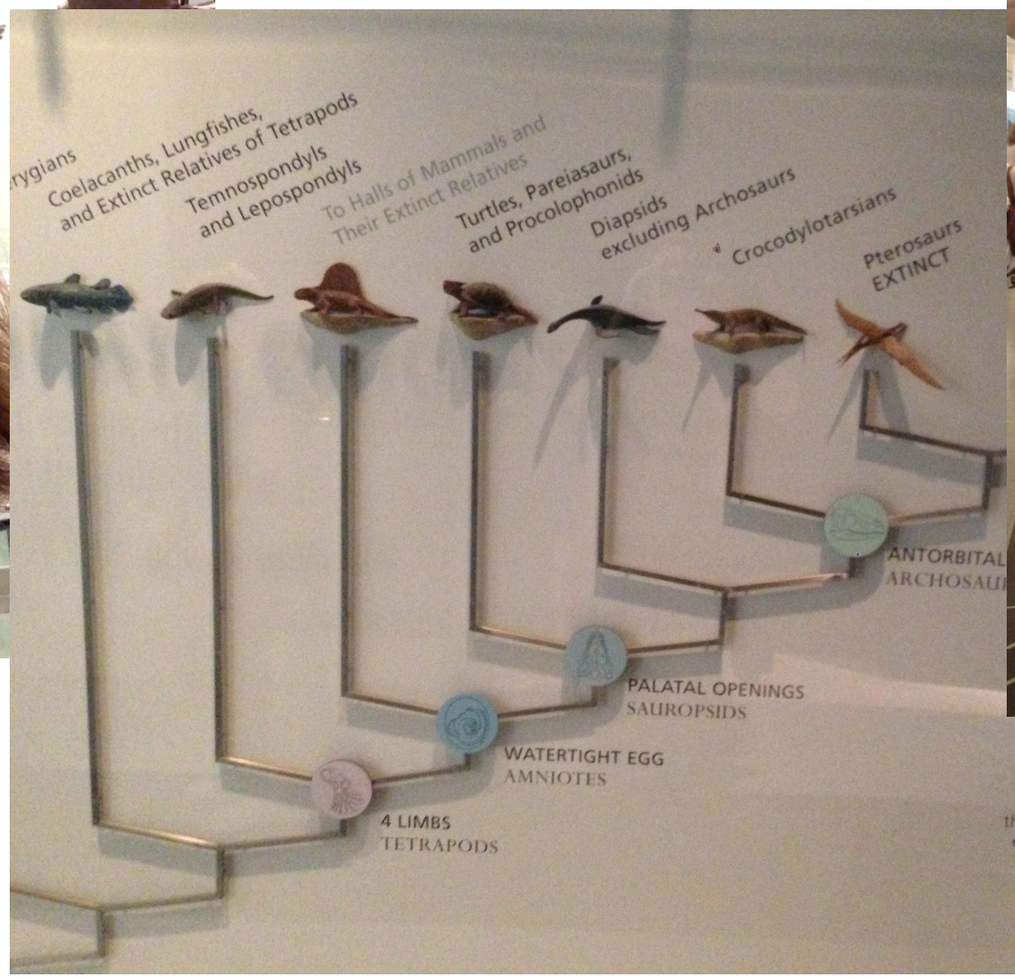
B. Is the Tree of Life even a tree?



Part I

Pre-genomics era

A walk down the Tree of Life



Compatible splits

Definition

An *X-split* $A|B$ is a bipartition of X into non-empty subsets A, B.

Definition

A pair of X-splits $A_1|B_1$ and $A_2|B_2$ is *compatible* if at least one of the sets $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$, or $B_1 \cap B_2$ is the empty set.

Theorem (Splits-equivalence theorem; Buneman (1971))

A set of X-splits is induced by an X-tree iff it is compatible.

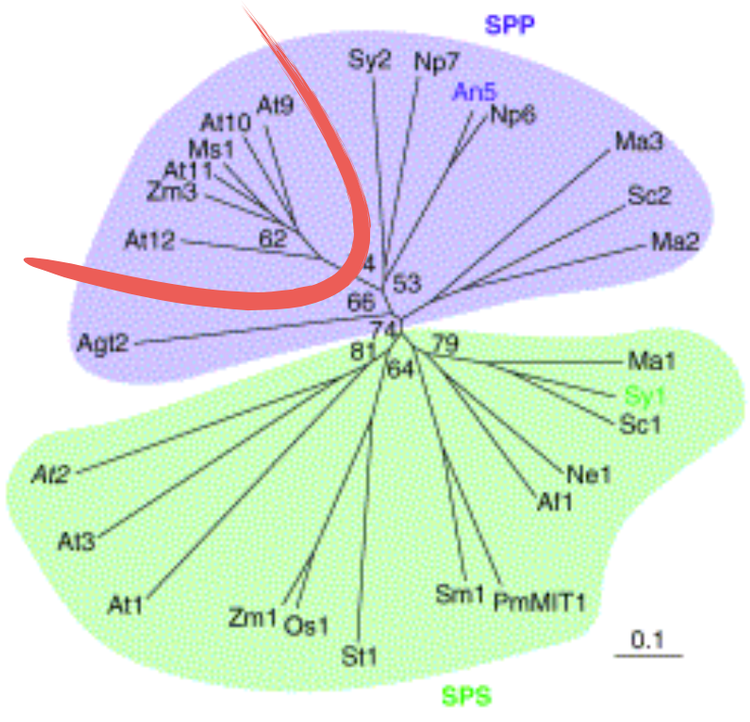
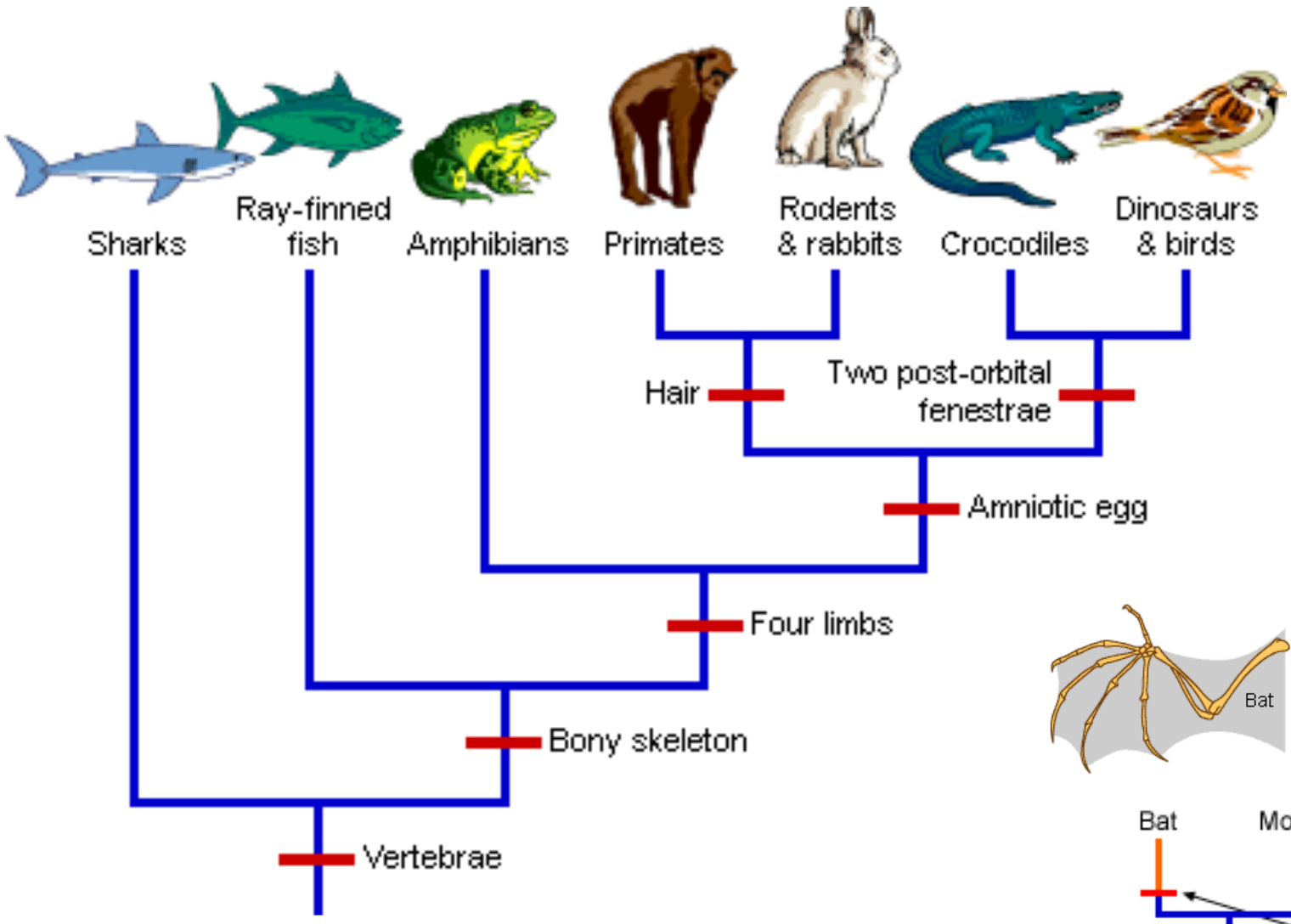
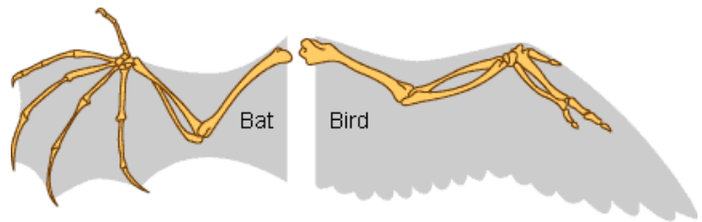
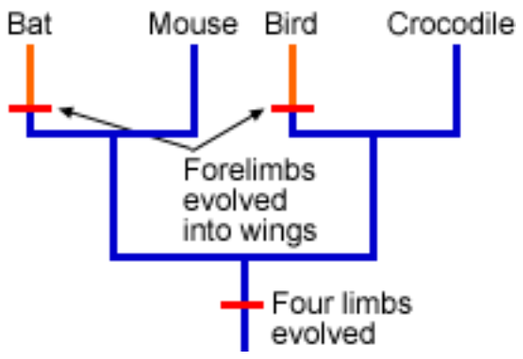


Figure from Salerno, Curatti, Trends Plant Sci. 2003 Feb;8(2):63-9.

Synapomorphies & homoplasies



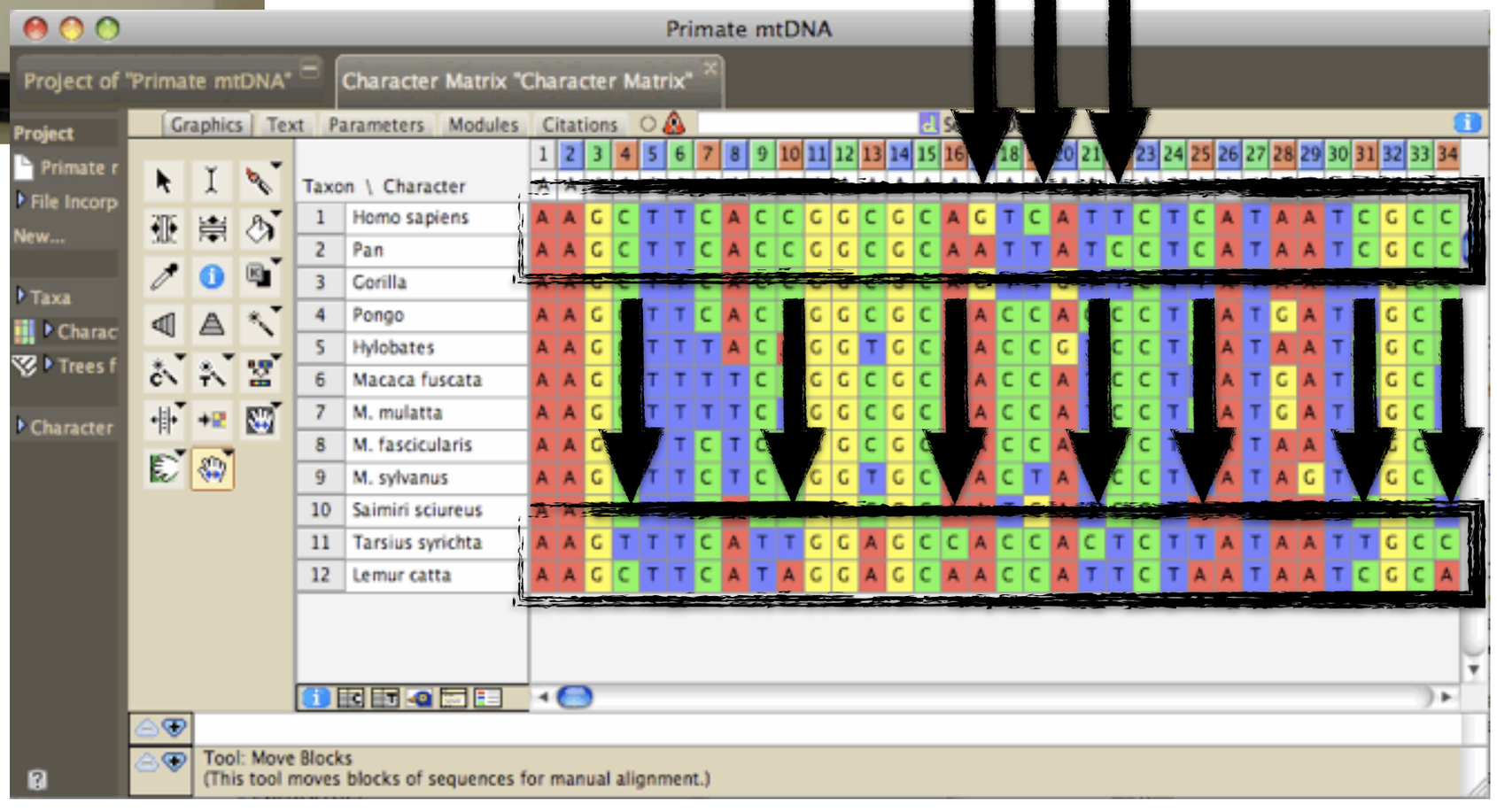




parsimony

Figures by University of California Museum of Paleontology's Understanding Evolution

Molecular systematics



Snapshot from Mesquite

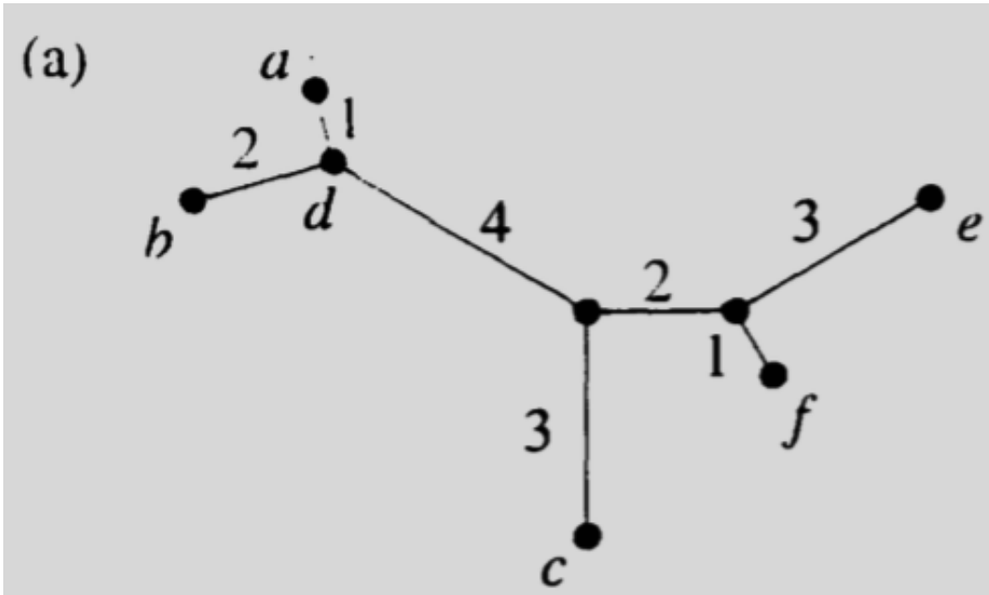
Tree metrics

Definition

A function $\delta : X \times X \to \mathbb{R}$ is a *tree metric* if there is an X-tree $\mathcal{T} = (T; \phi)$ and a weighting $w : E(T) \to \mathbb{R}_+$ such that for all x, y

$$\delta(x, y) = d_{(\mathcal{T};w)}(x, y) := \sum_{e \in P(\mathcal{T};x,y)} w(e),$$

where $P(\mathcal{T}; x, y)$ is the unique path between $\phi(x)$ and $\phi(y)$.
The *tree metric representation* $(\mathcal{T}; w)$ of δ is unique.

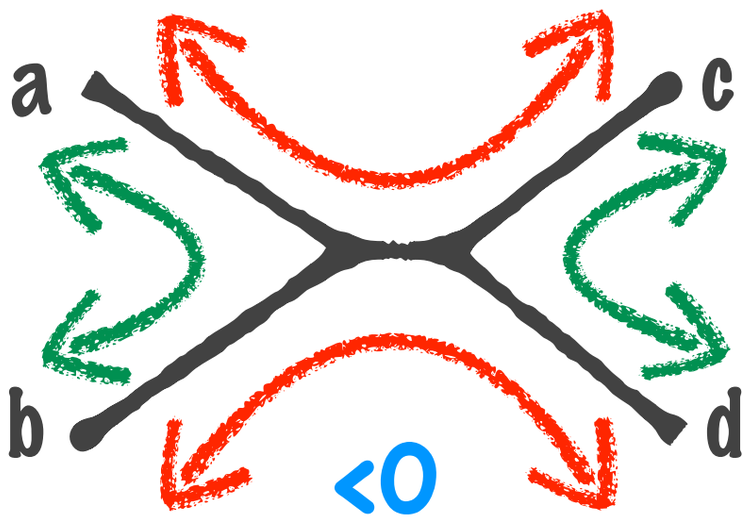


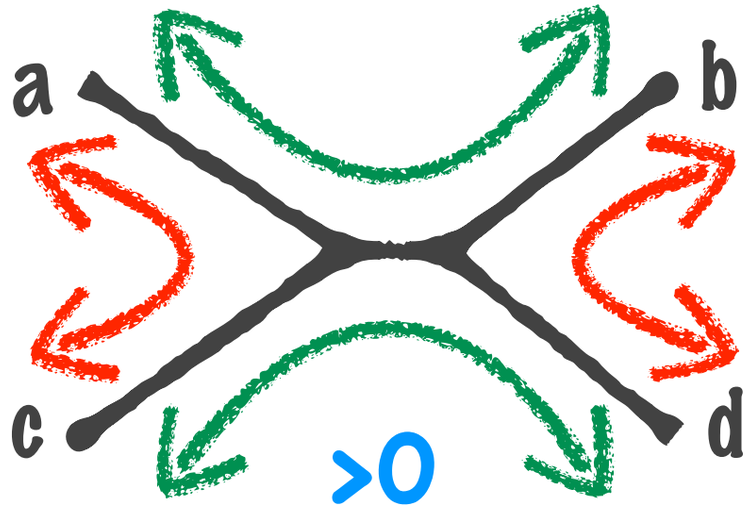
(b)

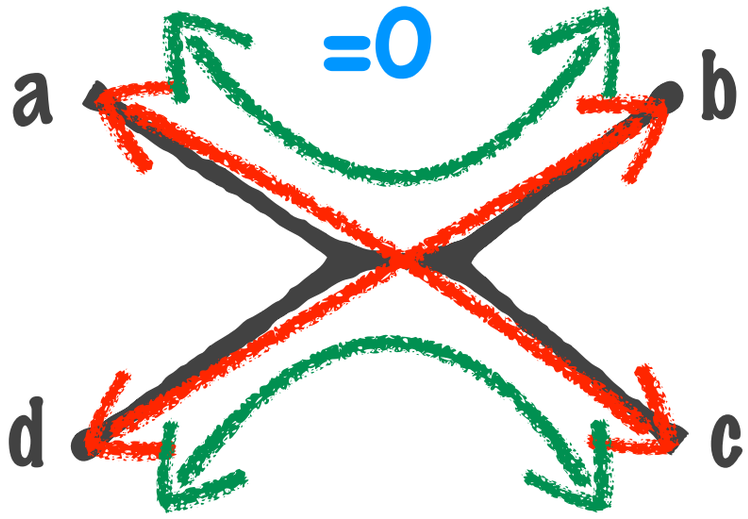
	a	b	c	d	e	f
a	0	3	8	1	10	8
b		0	9	2	11	9
c			0	7	8	6
d				0	9	7
e					0	4
f						0

Figure from Semple & Steel, Phylogenetics (2003).

Which quartet topology?

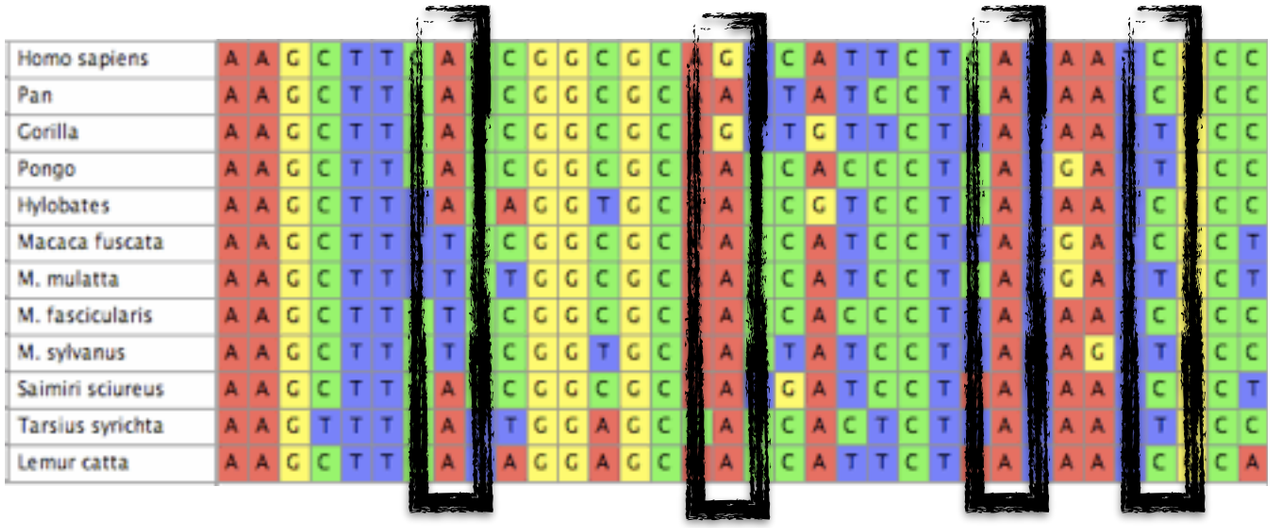




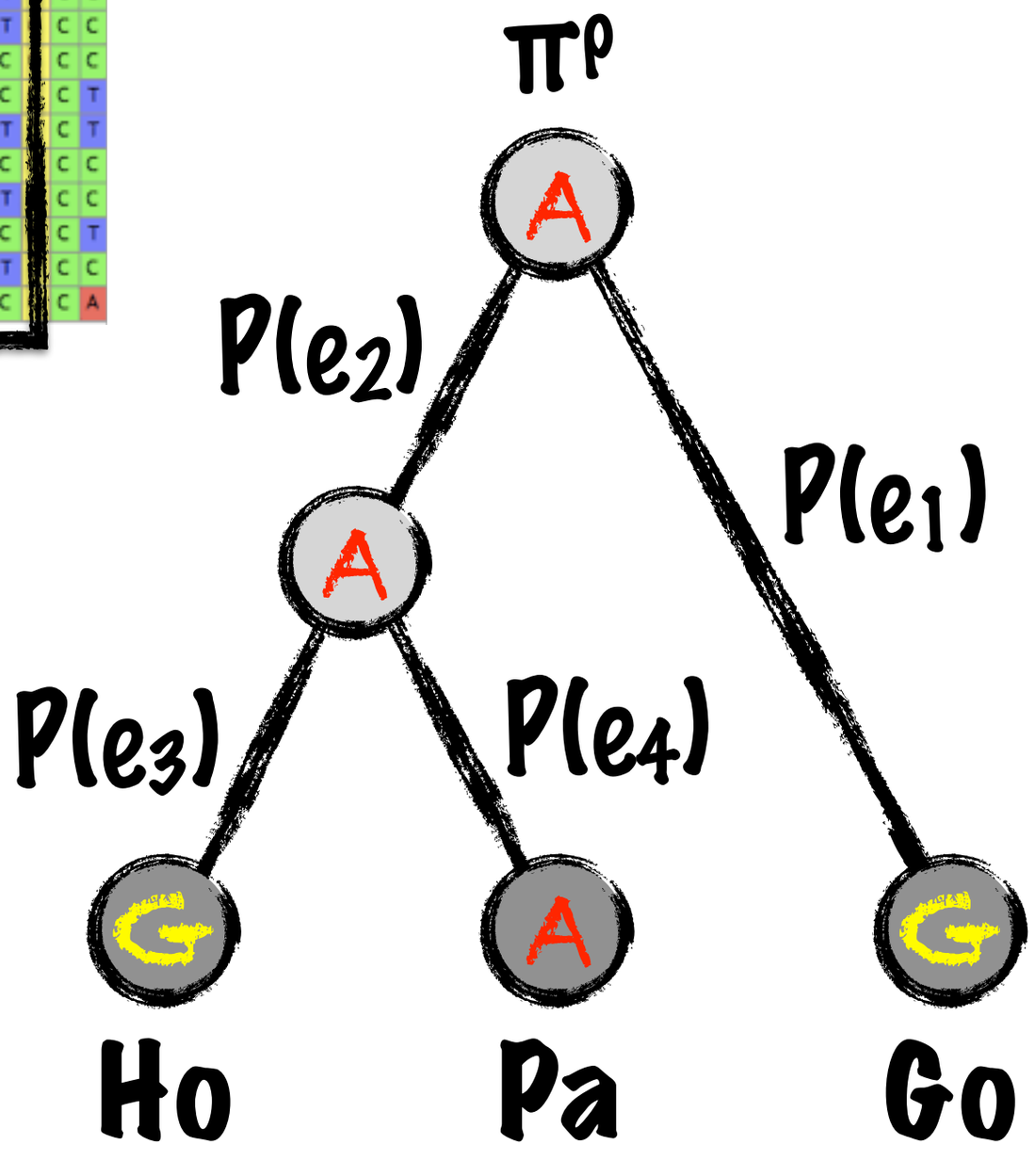


$$\delta(a,b) + \delta(c,d) - \delta(a,c) - \delta(b,d)$$

Markov process on a tree



k columns
=
k i.i.d. samples

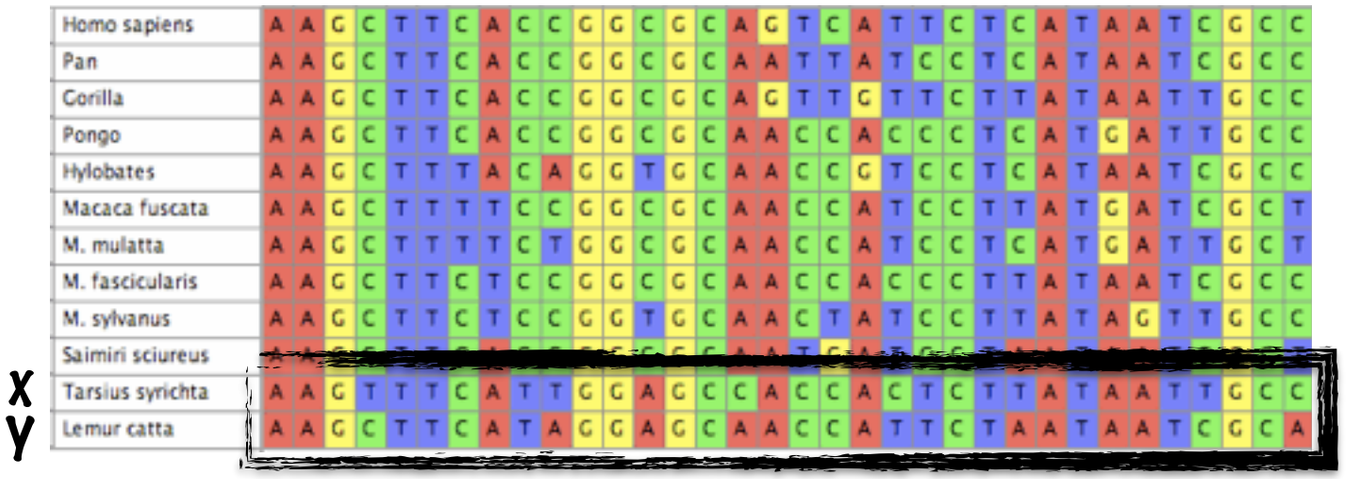


Back to tree metrics

Definition

Let F^{xy} be the matrix whose entries correspond to the joint distribution at the leaves x and y. The *log-det distance* is

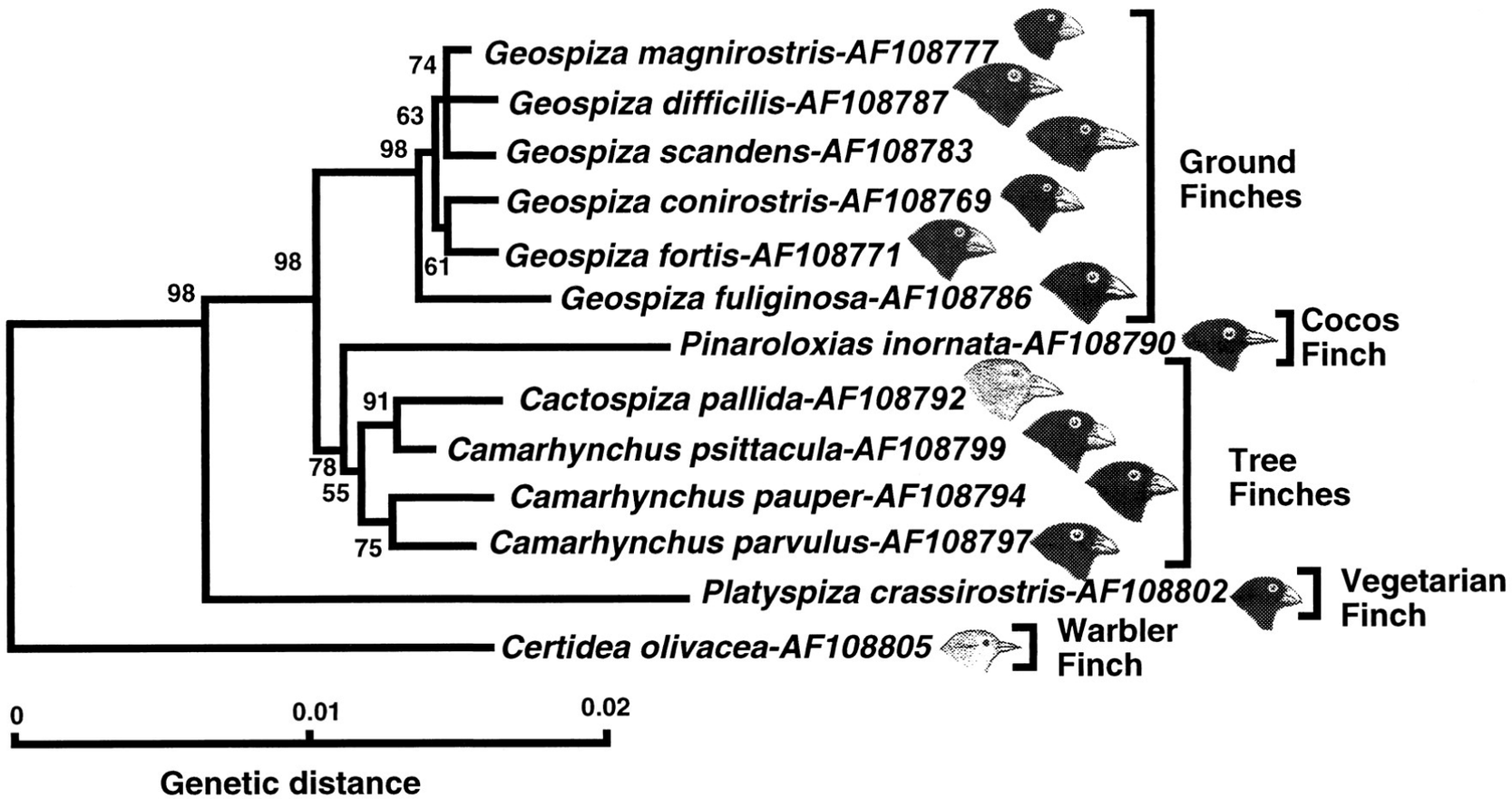
$$\delta(x, y) = -\log(\det(F^{xy})).$$



Theorem (Steel (1994))

Assume $\pi^\rho > 0$ *and* $|\det P(e)| \neq 0, 1$ *for all e. Then the log-det distance is a tree metric with corresponding X-tree* $\mathcal{T}$*.*

Back to Darwin's finches



Neighbor-Joining tree of combined cytb and cr sequences.
(From: Akie Sato et al. PNAS 1999;96:5101-5106)

Identifiability

The distribution of a "column" is given by:

$$p_\chi^{\mathcal{T}}(\theta) := \sum_{\substack{\bar{\chi}: V(T)\to C \\ \bar{\chi}\circ\phi=\chi}} \pi^{\rho}_{\bar{\chi}(\rho)} \prod_{e=(u,v)\in E(T)} P(e)_{\bar{\chi}(u),\bar{\chi}(v)}.$$

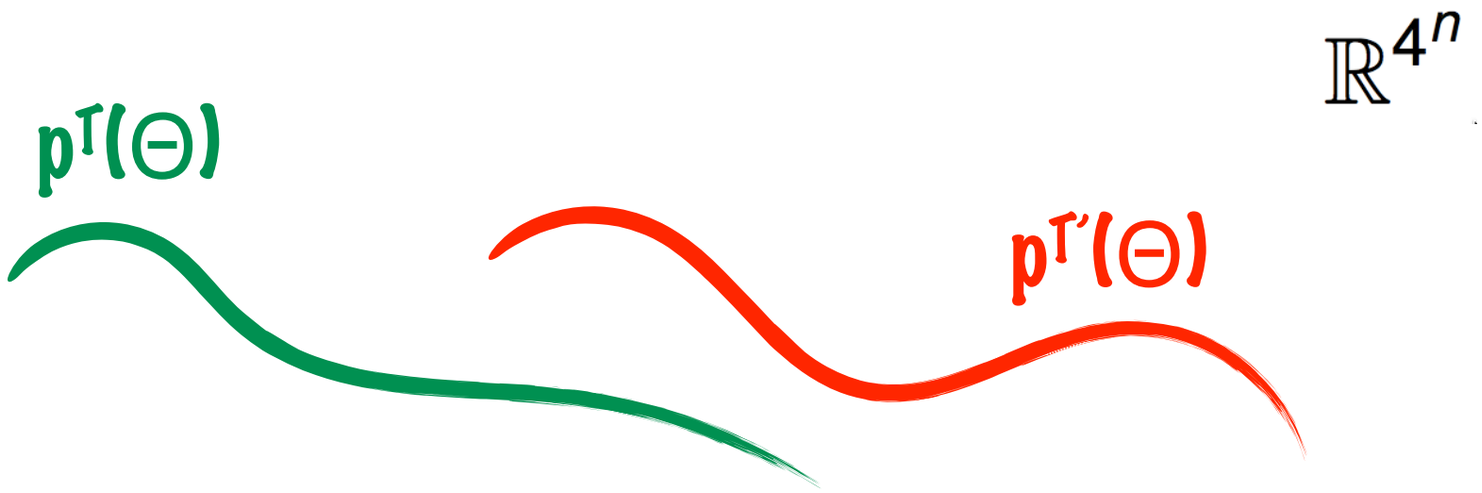
Definition

The tree is *identifiable* if $\mathcal{T} \neq \mathcal{T}'$ implies $p^{\mathcal{T}}(\theta) \neq p^{\mathcal{T}'}(\theta')$.

Theorem (Steel (1994))

If $\pi^\rho > 0$ *and* $|\det P(e)| \neq 0, 1$, *the tree is* identifiable.

Identifiability



Definition

The tree is *identifiable* if $\mathcal{T} \neq \mathcal{T}'$ implies $p^{\mathcal{T}}(\theta) \neq p^{\mathcal{T}'}(\theta')$.

Theorem (Steel (1994))

If $\pi^{\rho} > 0$ *and* $|\det P(e)| \neq 0, 1$*, the tree is* identifiable*.*

Likelihood-based inference

Definition

Given sequences of length k, i.e., $(\chi^i)_{i=1}^k$, the maximum likelihood estimator (MLE) is

$$\hat{\mathcal{T}}_k \in \arg\max \left\{ \prod_{i=1}^{k} p_{\chi^i}^{\mathcal{T}}(\theta) \ : \ \mathcal{T}, \theta \in \Theta \right\}.$$

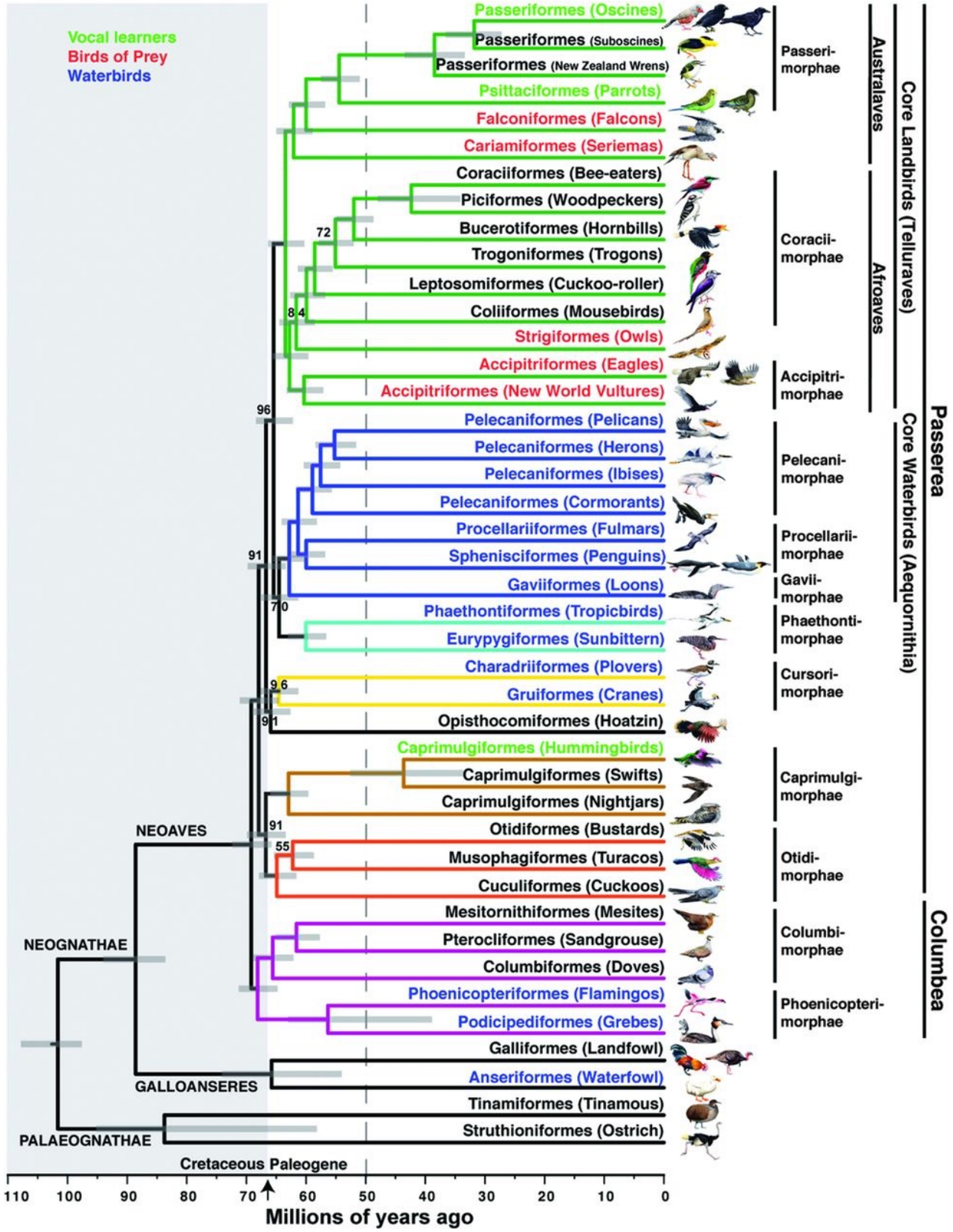
Theorem (Chang (1996))

The MLE is consistent, *i.e.,* $\hat{\mathcal{T}}_k \to \mathcal{T}$ *as* $k \to +\infty$.

Theorem (Chor & Tuller (2006); R. (2006))

Computing the MLE is NP-hard.

How much data is needed?

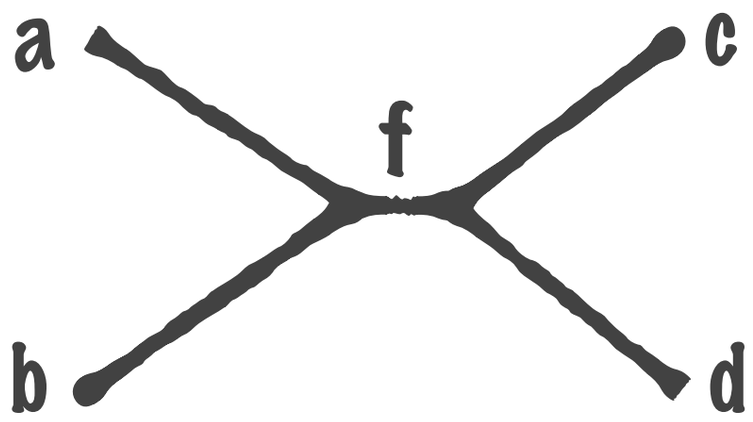


Genome-scale phylogeny of birds. (From: Erich D. Jarvis et al. Science 2014;346:1320-1331)

Short branches

Theorem (Steel & Székely (2002))

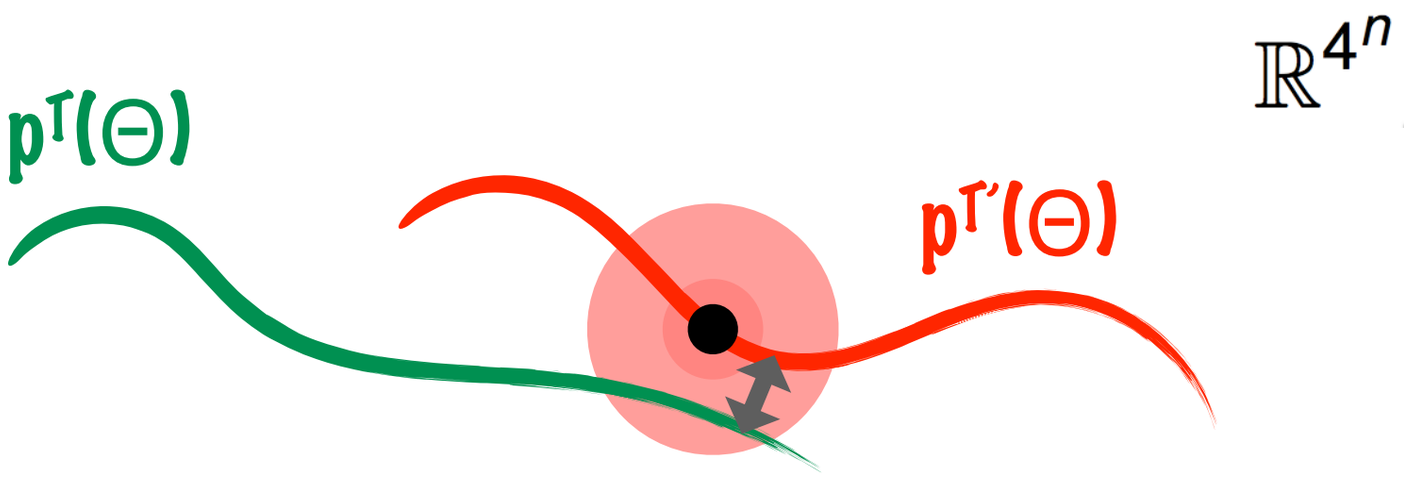
Under the symmetric 2-state Markov model on 4 *species, reconstructing the phylogeny with high probability requires* $k \geq Cf^{-2}$ *sites, where* f *is the length of the internal branch.*



Short branches

Theorem (Steel & Székely (2002))

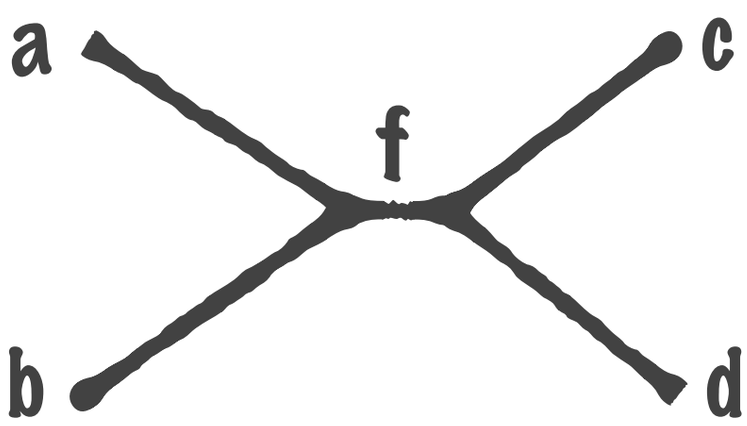
Under the symmetric 2-state Markov model on 4 *species, reconstructing the phylogeny with high probability requires* $k \geq Cf^{-2}$ *sites, where* f *is the length of the internal branch.*

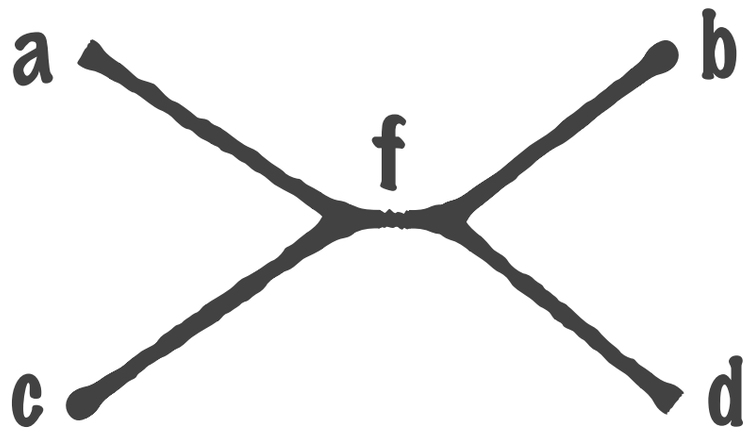


Short branches

Theorem (Steel & Székely (2002))

Under the symmetric 2-state Markov model on 4 *species, reconstructing the phylogeny with high probability requires* $k \geq Cf^{-2}$ *sites, where* f *is the length of the internal branch.*





Short branches

Theorem (Steel & Székely (2002))

Under the symmetric 2-state Markov model on 4 *species, reconstructing the phylogeny with high probability requires* $k \geq Cf^{-2}$ *sites, where* f *is the length of the internal branch.*

Total variation distance

- For two discrete measures Q = {q_i}_i and Q'={q'_i}

$$\|Q - Q'\|_{\mathrm{TV}} = \sup_A |Q(A) - Q'(A)|$$

- 1-TV = sum of Type I and Type II errors for likelihood ratio test

Short branches

Theorem (Steel & Székely (2002))

Under the symmetric 2-state Markov model on 4 *species, reconstructing the phylogeny with high probability requires* $k \geq Cf^{-2}$ *sites, where* f *is the length of the internal branch.*

Hellinger distance

- Under the same setting

$$H^2(Q, Q') = \sum \left(\sqrt{q_i} - \sqrt{q'_i} \right)^2$$

- Factorizes nicely

$$\frac{1}{2} H^2(Q^{\otimes k}, Q'^{\otimes k}) = 1 - \left(1 - \frac{1}{2} H^2(Q, Q') \right)^k$$

- Moreover

$$\|Q - Q'\|_{\mathrm{TV}} \leq H(Q, Q')$$

Depth

A special case of a more general phenomenon:

Theorem (Mossel (2004))

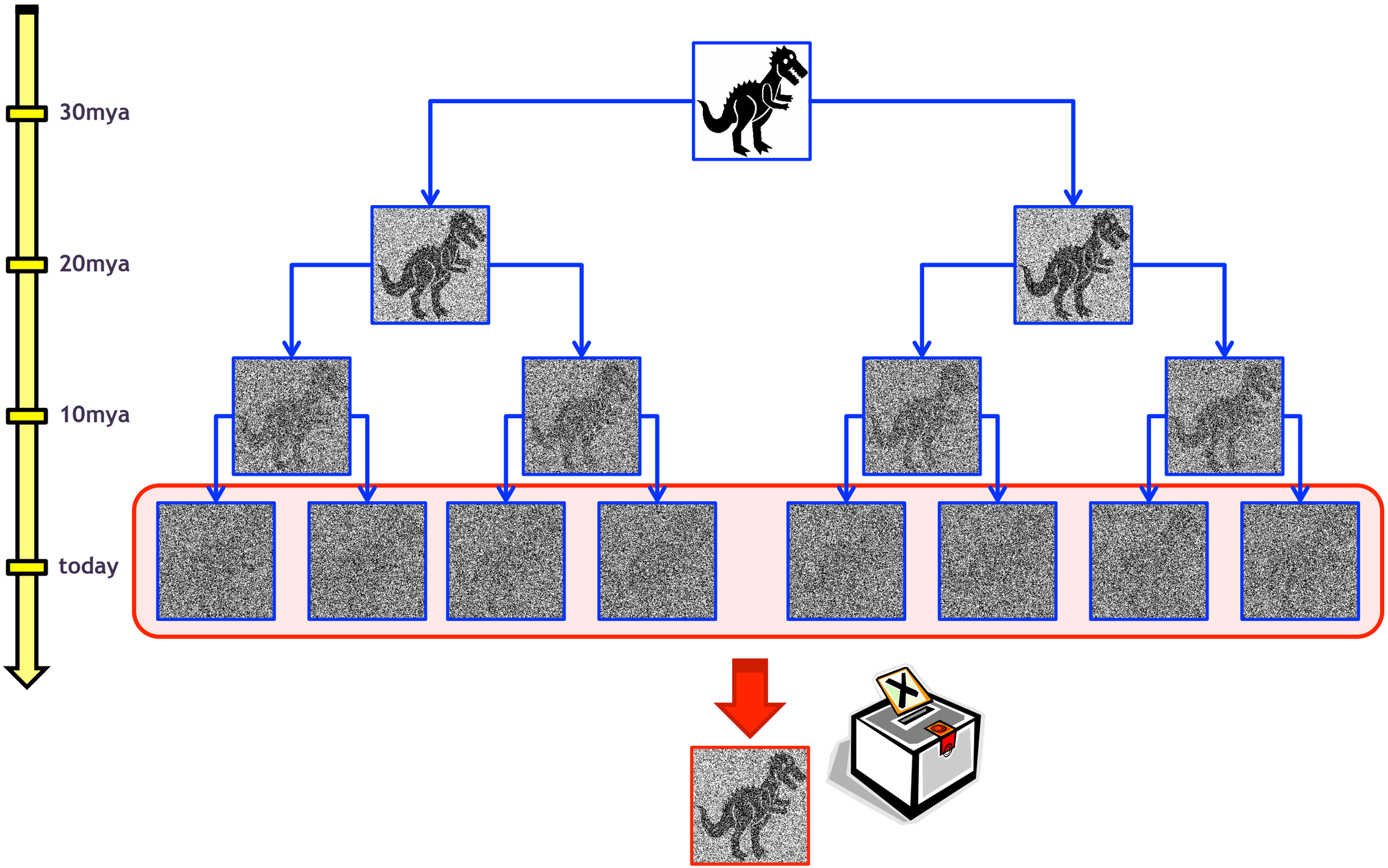
Under the symmetric 2-state Markov model on n species with branches of weight f, reconstructing the phylogeny with high probability from k sites requires in general

$$k = \begin{cases} \Theta(f^{-2}\log n), & \text{if } f < f^*, \\ n^{\Theta(f)}, & \text{if } f \geq f^*, \end{cases}$$

for some critical f^*.

Matched for MLE (R. & Sly (2017)) and some tree metric-based methods (R. (2010)). In contrast other popular methods, such as Neighbor-Joining, may require exponentially (in n) more data (Lacey & Chang (2006)).

Correlation decay



Markov chain on two states

Observation (MC on line)

Let $(X_\tau)_{\tau=0}^{+\infty}$ *be MC on* $\{-1,+1\}$ *with transition*

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}, \qquad \text{where } p \in (0,1/2).$$

It converges to uniform distribution. Since $(-1,1)$ *is eigenvector with eigenvalue* $\theta = 1-2p \in (0,1)$*:*

$$\mathbb{E}[X_h \mid X_0] = \theta^h X_0$$

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

L_h

$X_0 = +1$

+1 +1 +1 -1 -1 -1 -1 +1 +1 +1 -1 +1 -1 +1 -1 +1 ...

$$\overline{X}_h = \frac{1}{|L_h|} \sum_{\tau \in L_h} X_\tau$$

Markov chain on a tree

Observation (MC on tree)

Let $(X_\tau)_{\tau \in \mathcal{T}}$ *be MC on complete binary tree* $\mathcal{T}$ *with transition*

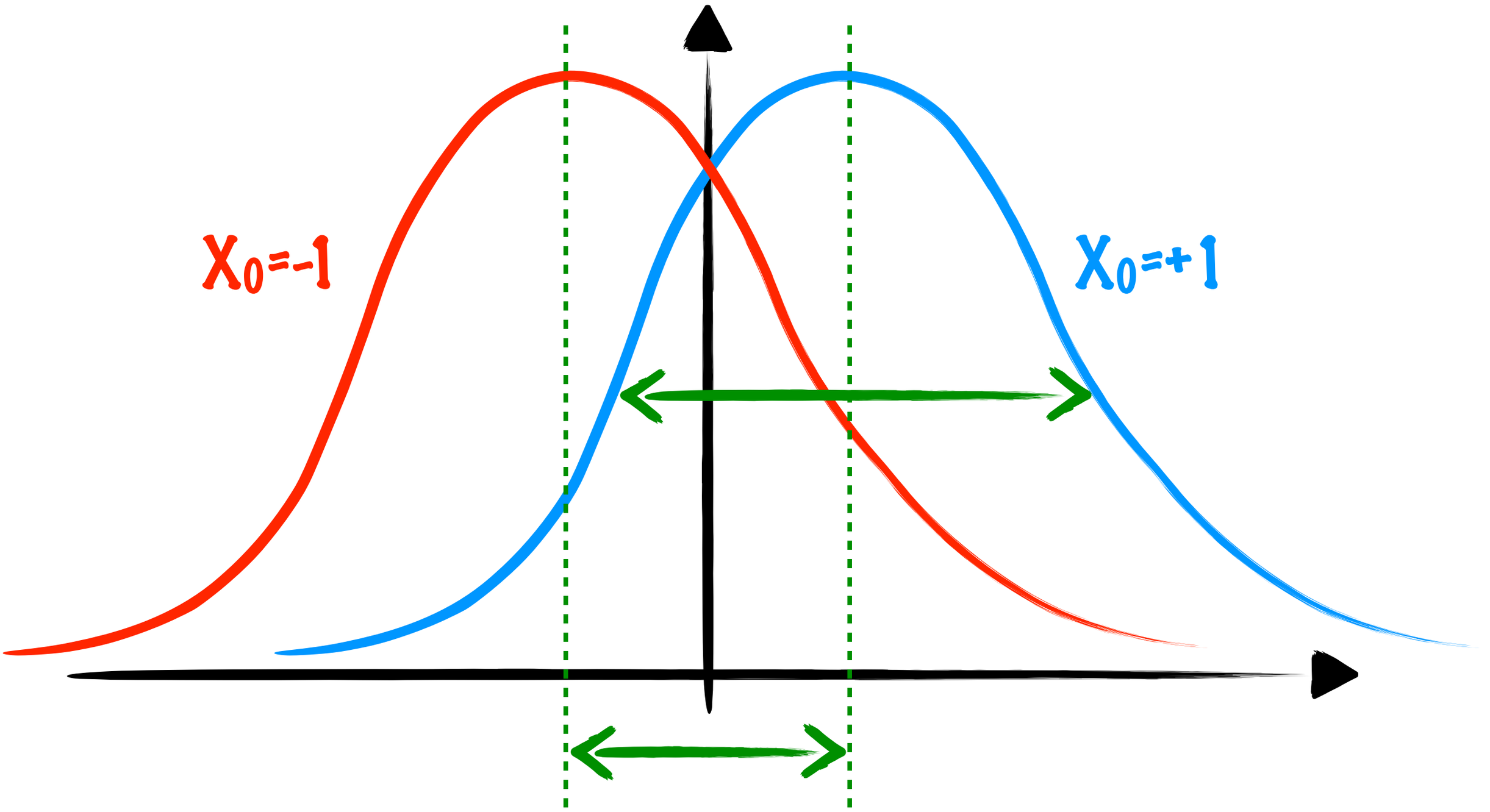
$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \text{ on } \{-1,+1\} \text{ where } p \in (0,1/2).$$

Letting $\overline{X}_h = \frac{1}{|L_h|} \sum_{\tau \in L_h} X_\tau$ *be average on* L_h *(i.e., level h):*

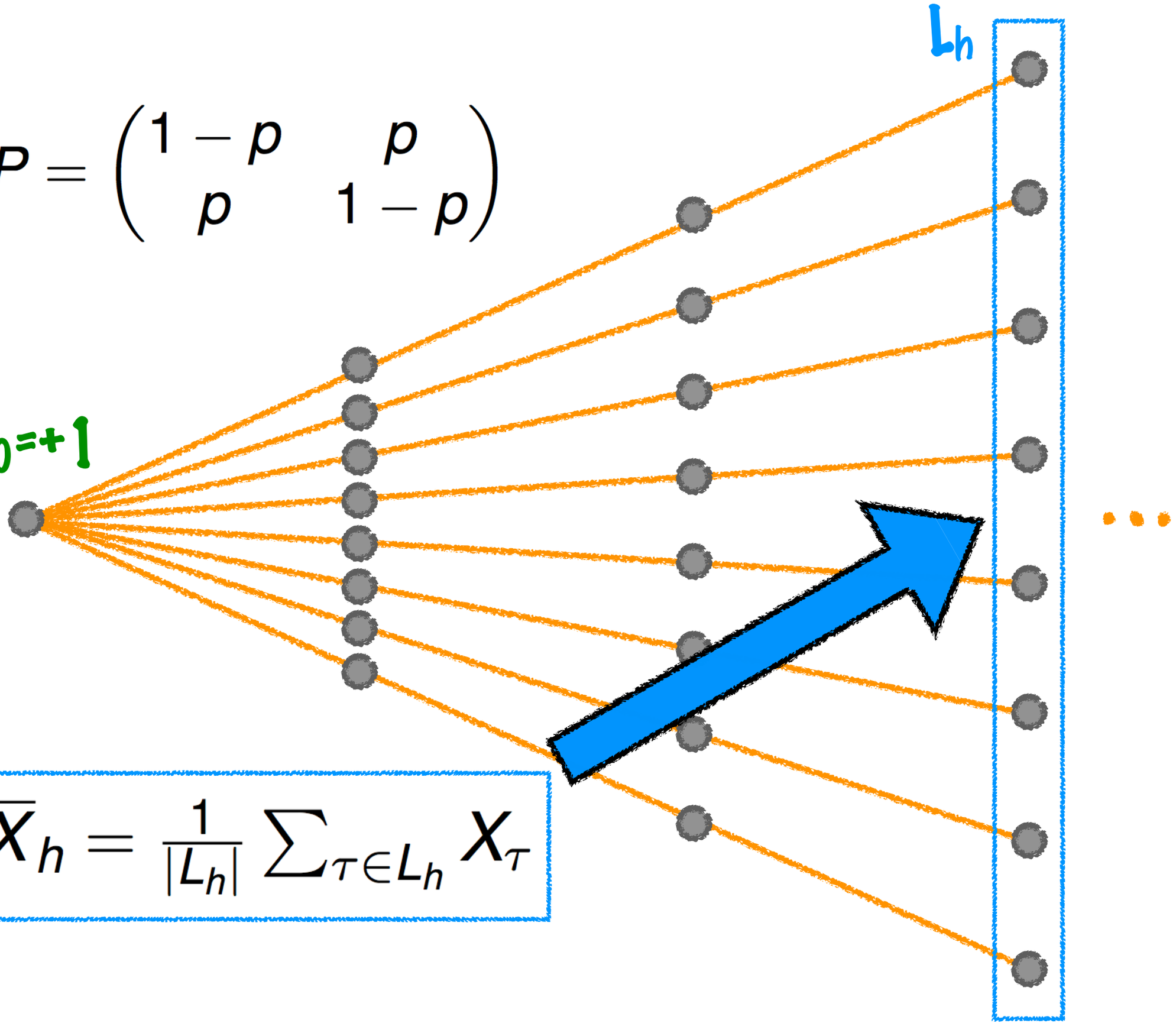
$$\mathbb{E}[\overline{X}_h \mid X_0] = \theta^h X_0$$

$$\mathrm{Var}[\overline{X}_h \mid X_0] = ?$$

Markov chain on a tree



$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$



$$\overline{X}_h = \frac{1}{|L_h|} \sum_{\tau \in L_h} X_\tau$$

Back-of-the-envelope: ignoring correlations

Observation (MC on star)

Let $(X_\tau)_{\tau\in\mathcal{S}}$ be MC on h-level star $\mathcal{S}$ with 2^h prongs & transition

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \text{ on } \{-1,1\} \text{ where } p \in (0,1/2).$$

Letting $\overline{X}_h = \frac{1}{2^h}\sum_{\tau\in L_h} X_\tau$ be average on L_h (i.e., level h):

$$\mathbb{E}[\overline{X}_h \mid X_0] = \theta^h X_0 \qquad \theta = 1-2p \in (0,1)$$

$$\mathrm{Var}[\overline{X}_h \mid X_0] = \frac{1}{2^h}\left[\mathbb{E}[X_\tau^2 \mid X_0] - \mathbb{E}[X_\tau \mid X_0]^2\right] = \frac{1}{2^h}[1-\theta^{2h}]$$

$$\frac{\left|\mathbb{E}[\overline{X}_h \mid X_0 = 1] - \mathbb{E}[\overline{X}_h \mid X_0 = -1]\right|}{\sqrt{\mathrm{Var}[\overline{X}_h \mid X_0 = 1]}} = 2\sqrt{\frac{(2\theta^2)^h}{1-\theta^{2h}}} \to \begin{cases} 0 & \text{if } 2\theta^2 < 1 \\ +\infty & \text{if } 2\theta^2 > 1 \end{cases}$$

Back to the tree

Observation (MC on tree)

Let $(X_\tau)_{\tau\in\mathcal{T}}$ *be MC on complete binary tree* $\mathcal{T}$ *with transition*

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \text{ on } \{-1,1\} \text{ where } p \in (0,1/2).$$

Letting $\overline{X}_h = |L_h|^{-1}\sum_{\tau\in L_h} X_\tau$ *be average on* L_h *(i.e., level h):*

$$\mathbb{E}[\overline{X}_h \,|\, X_0] = \theta^h X_0$$

$$\frac{\left|\mathbb{E}[\overline{X}_h \,|\, X_0 = 1] - \mathbb{E}[\overline{X}_h \,|\, X_0 = -1]\right|}{\sqrt{\mathrm{Var}[\overline{X}_h \,|\, X_0 = 1]}} \to \begin{cases} 0 & \text{if } 2\theta^2 < 1 \\ C & \text{if } 2\theta^2 > 1 \end{cases}$$

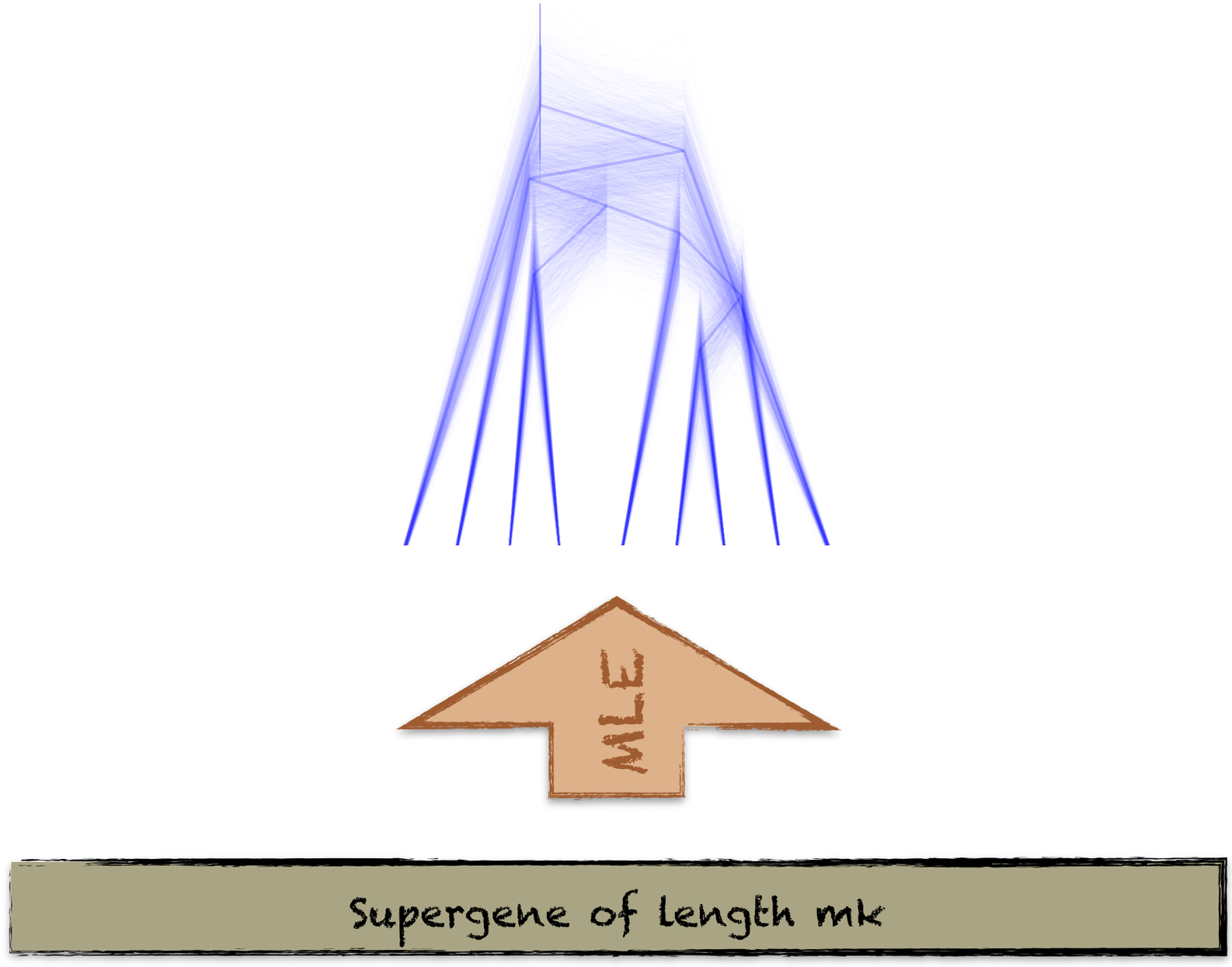
where $C > 0$ *is a constant.*

Note: $2\theta^2 = 1$ is the Kesten-Stigum threshold

Part II

More data, more problems

Easy: concatenate

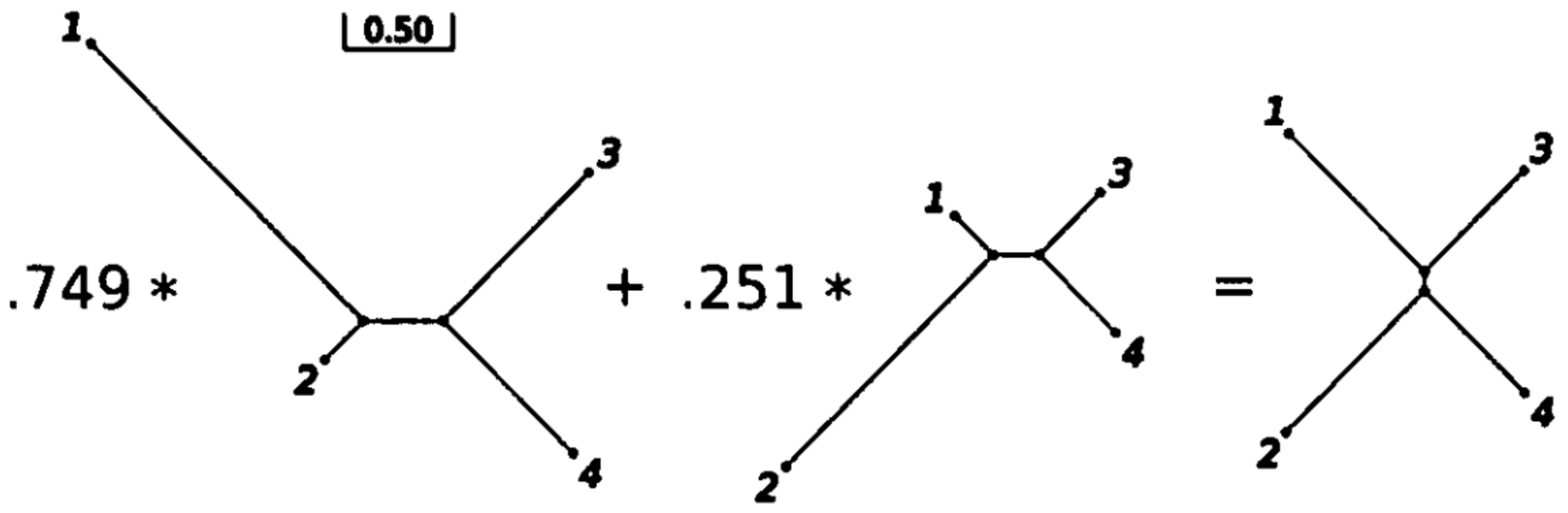


Mixed-up trees

Using algebraic geometry (Sturmfels & Sullivant, JCB (2005)):

Theorem (Matsen & Steel, SB (2007))

Phylogenetic mixtures on a single tree can mimic a tree of another topology.

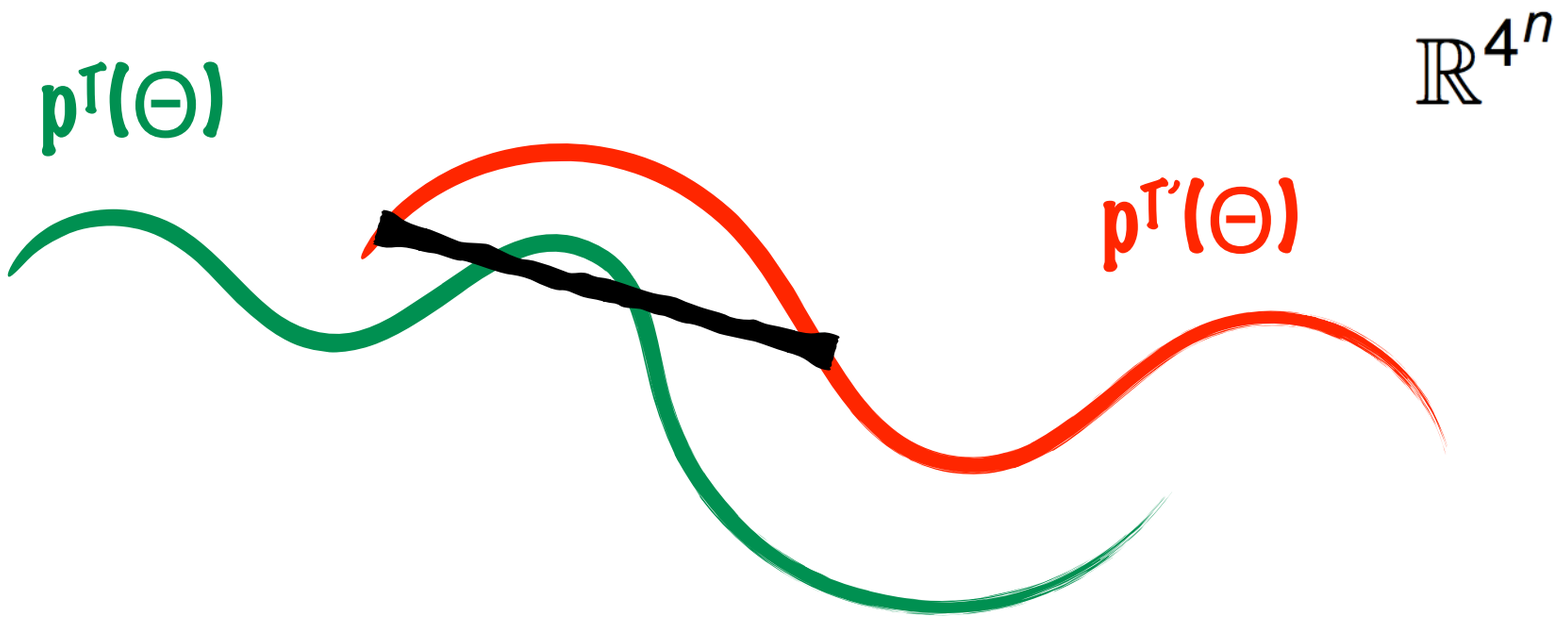


Mixed-up trees

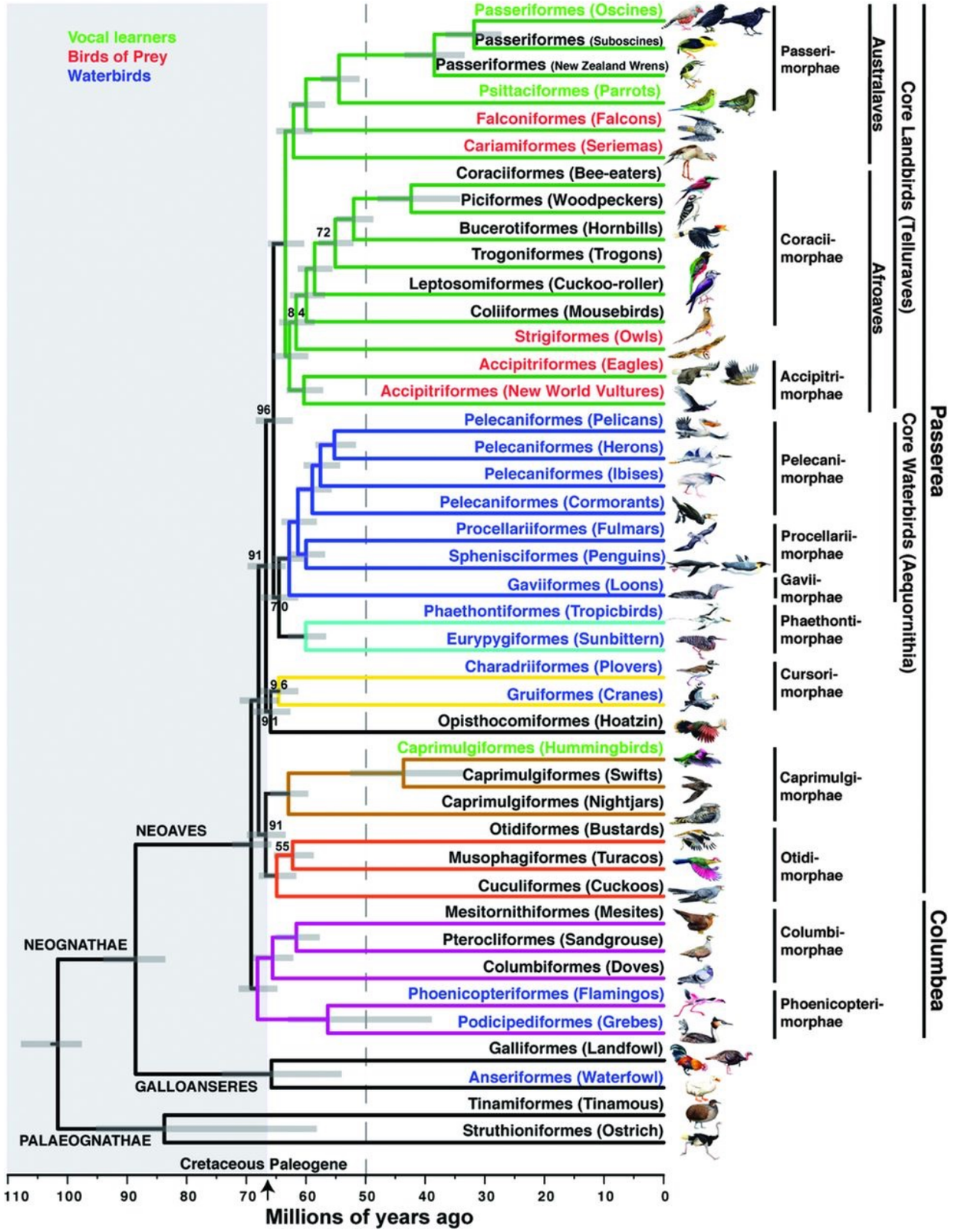
Using algebraic geometry (Sturmfels & Sullivant, JCB (2005)):

Theorem (Matsen & Steel, SB (2007))

Phylogenetic mixtures on a single tree can mimic a tree of another topology.

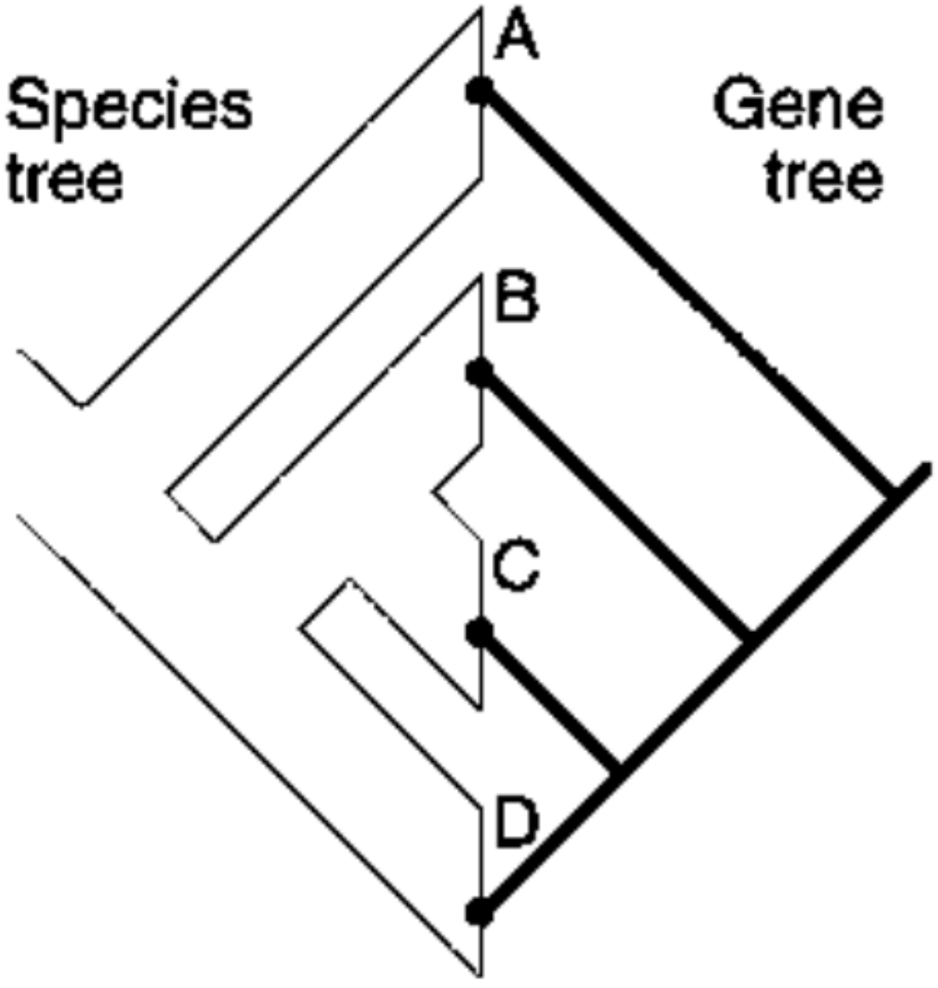


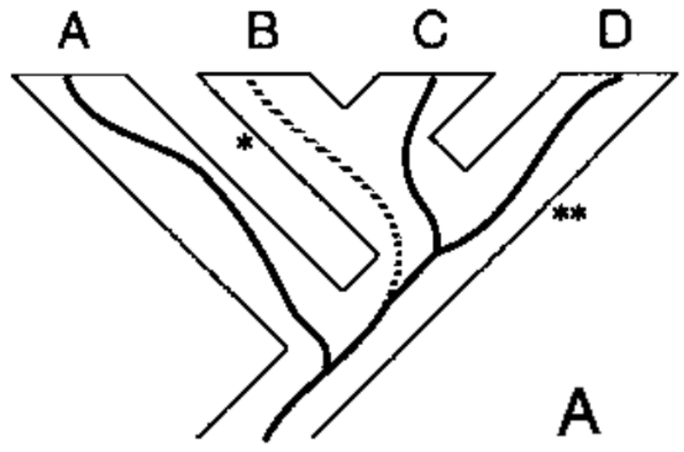
Back to those birds

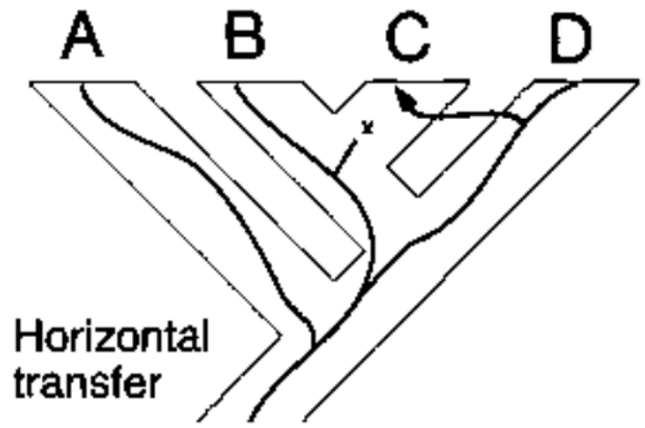


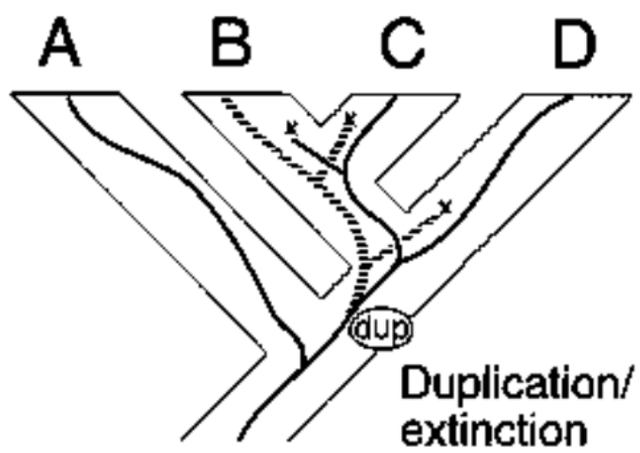
Genome-scale phylogeny of birds. (From: Erich D. Jarvis et al. Science 2014;346:1320-1331)

Species tree v. “gene” trees









Figures from Maddison, Syst. Biol. (1997).

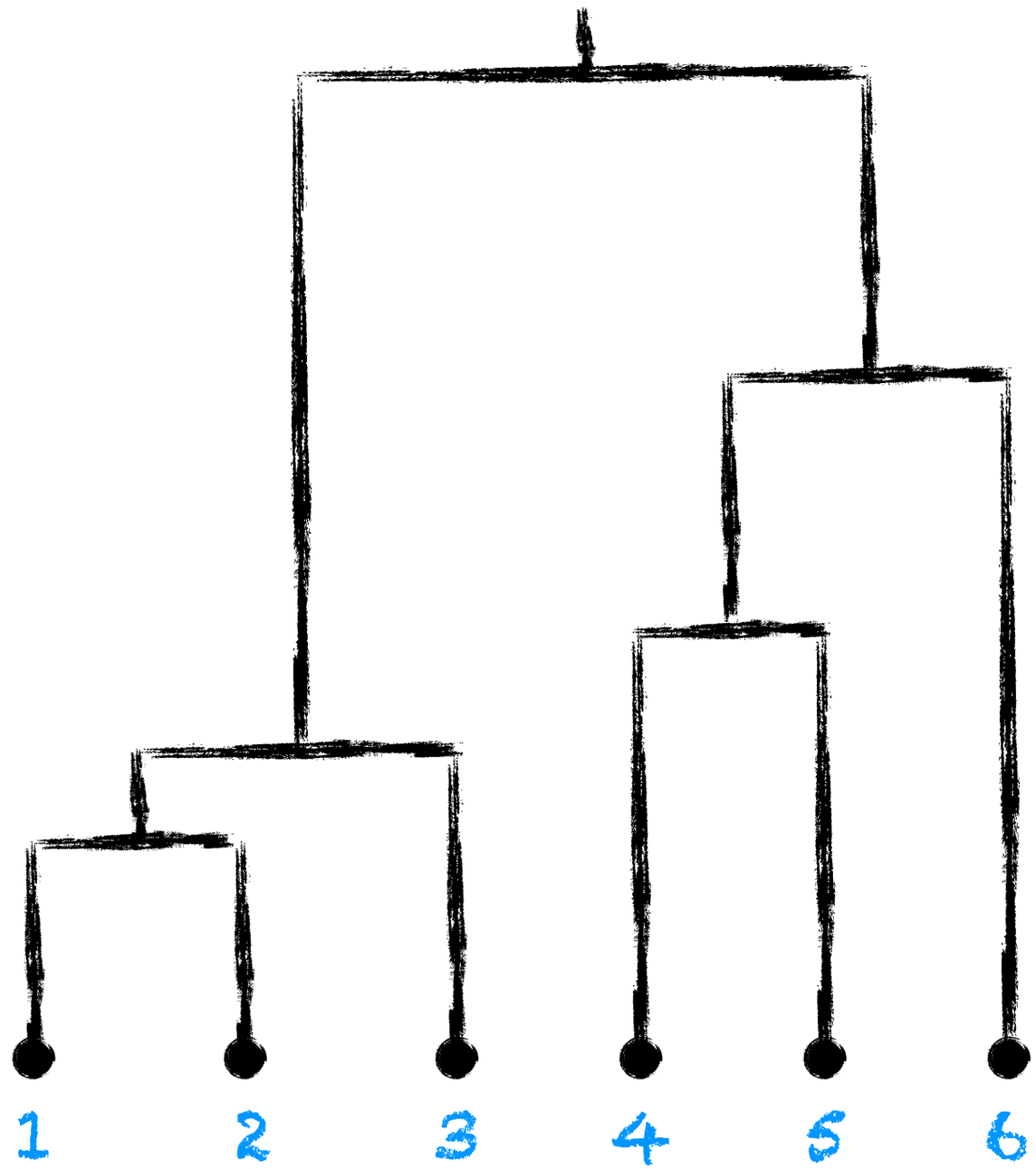
Coalescent processes

Kingman's coalescent

- Continuous-time process on partitions of [n]
- Start with {1},…,{n}
- Each pair of sets in the current partition independently merges at exponential rate 1
- Stop when {1,…,n} is reached

Application to population genetics

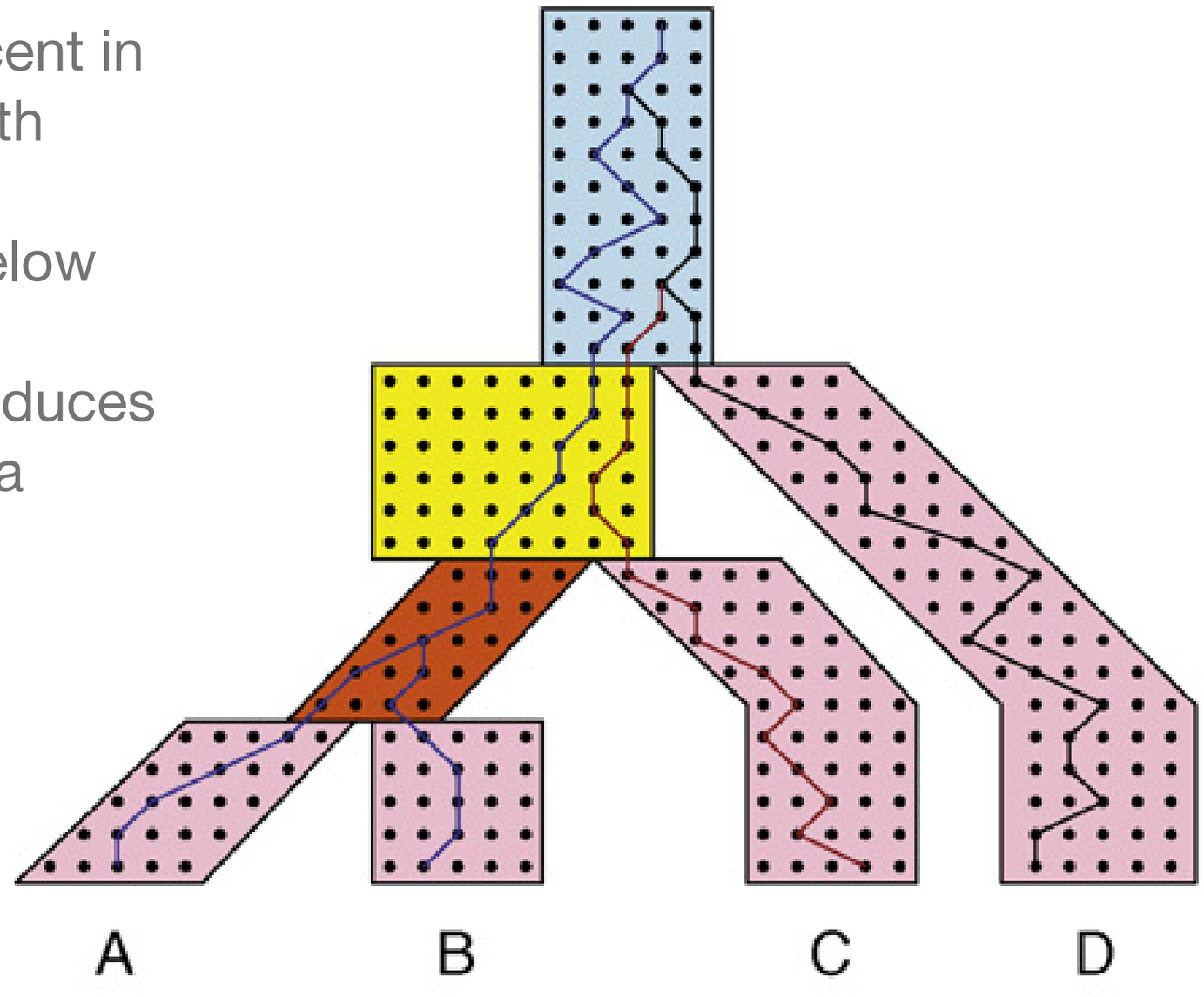
- Coalescent is commonly used to model lineages backwards in time in a population



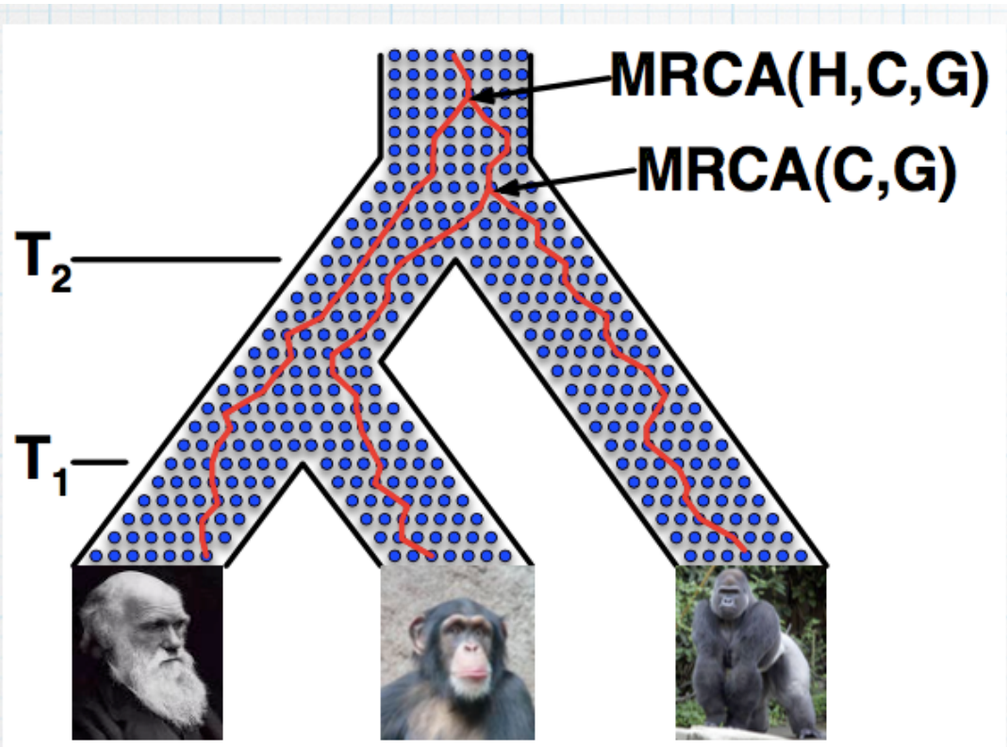
Coalescent processes

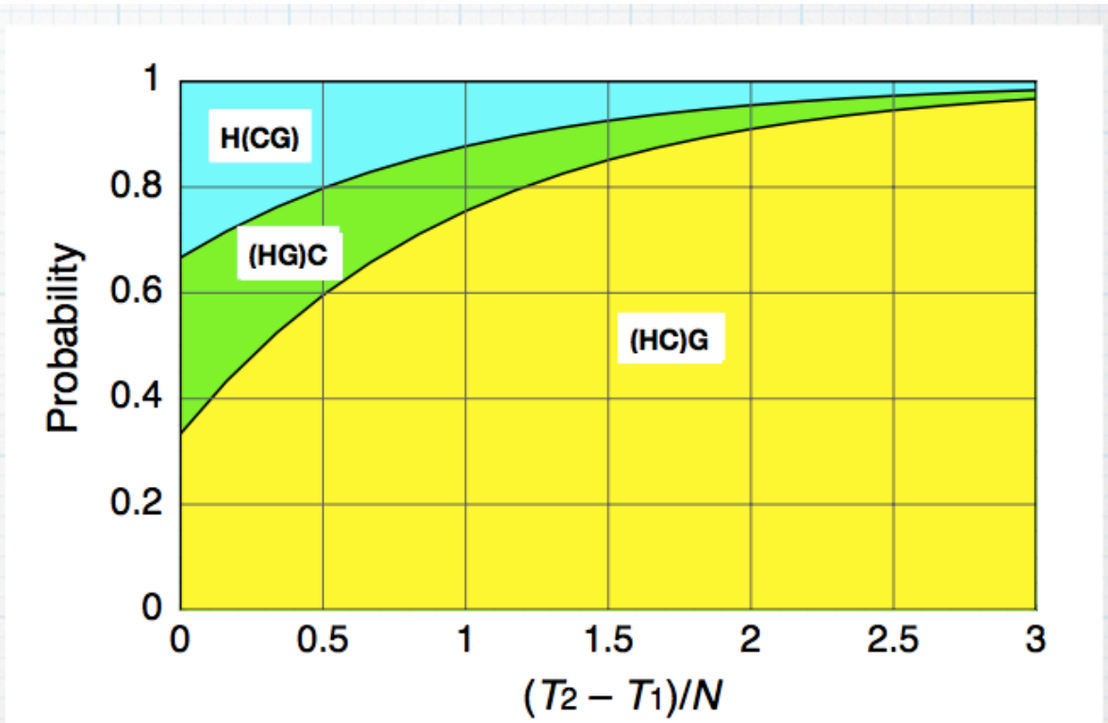
Multispecies coalescent

- S: species tree
- Perform Kingman's coalescent in each population starting with lineages entering from populations immediately below
- Stitching together the corresponding lineages produces a tree which we refer to as a **gene tree**



Coalescent processes





$$\mathbf{P}[((H,C),G)] = 1 - \frac{2}{3}e^{-(T_2-T_1)/N}$$

$$\mathbf{P}[((H,G),C)] = \frac{1}{3}e^{-(T_2-T_1)/N}$$

$$\mathbf{P}[(H,(C,G))] = \frac{1}{3}e^{-(T_2-T_1)/N}$$

Figures by Luay Nakhleh

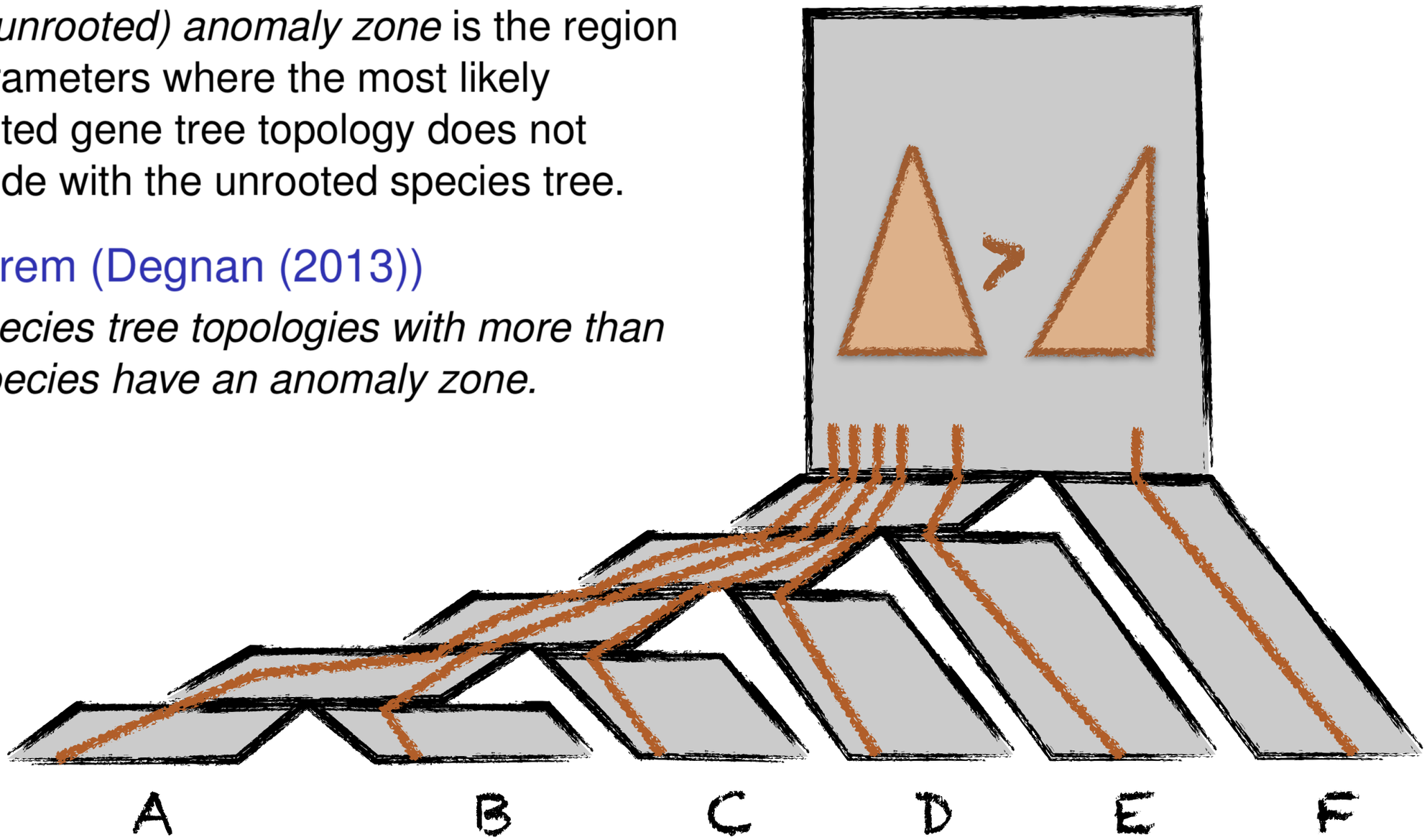
Anomaly zone

Definition

The *(unrooted) anomaly zone* is the region of parameters where the most likely unrooted gene tree topology does not coincide with the unrooted species tree.

Theorem (Degnan (2013))

All species tree topologies with more than six species have an anomaly zone.



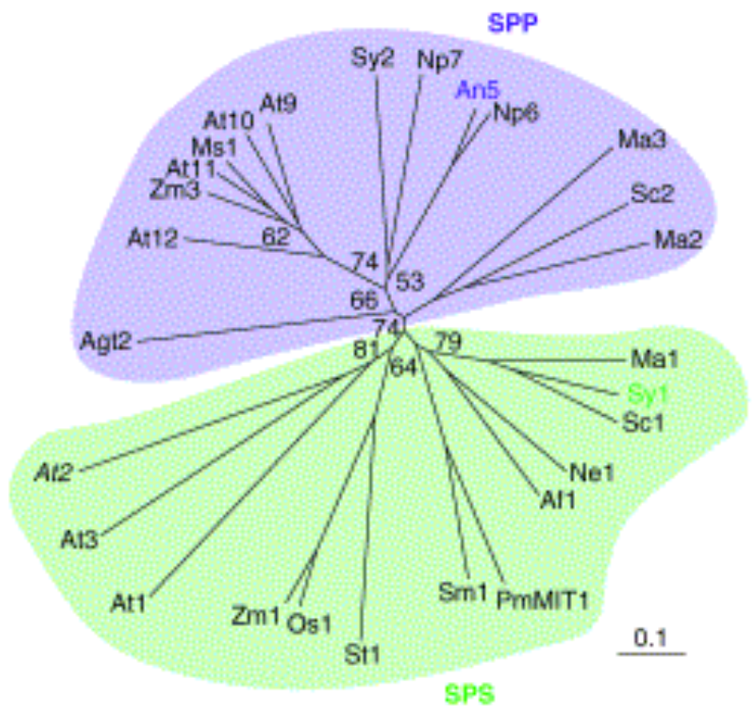
Rooted version: Degnan & Rosenberg (2006)

Split-based approaches

Theorem (Allman et al. (2018))

The unrooted species tree topology is identifiable from the split frequencies.

Proof is based on showing that the **internode distance**, i.e. the average over genes of the graph distance between pairs of species, is a tree metric consistent with the species tree. (Many other split-based methods are not consistent.)



Rooted version: Allman et al. (2011)

Figure from Salerno, Curatti, Trends Plant Sci. 2003 Feb;8(2):63-9.

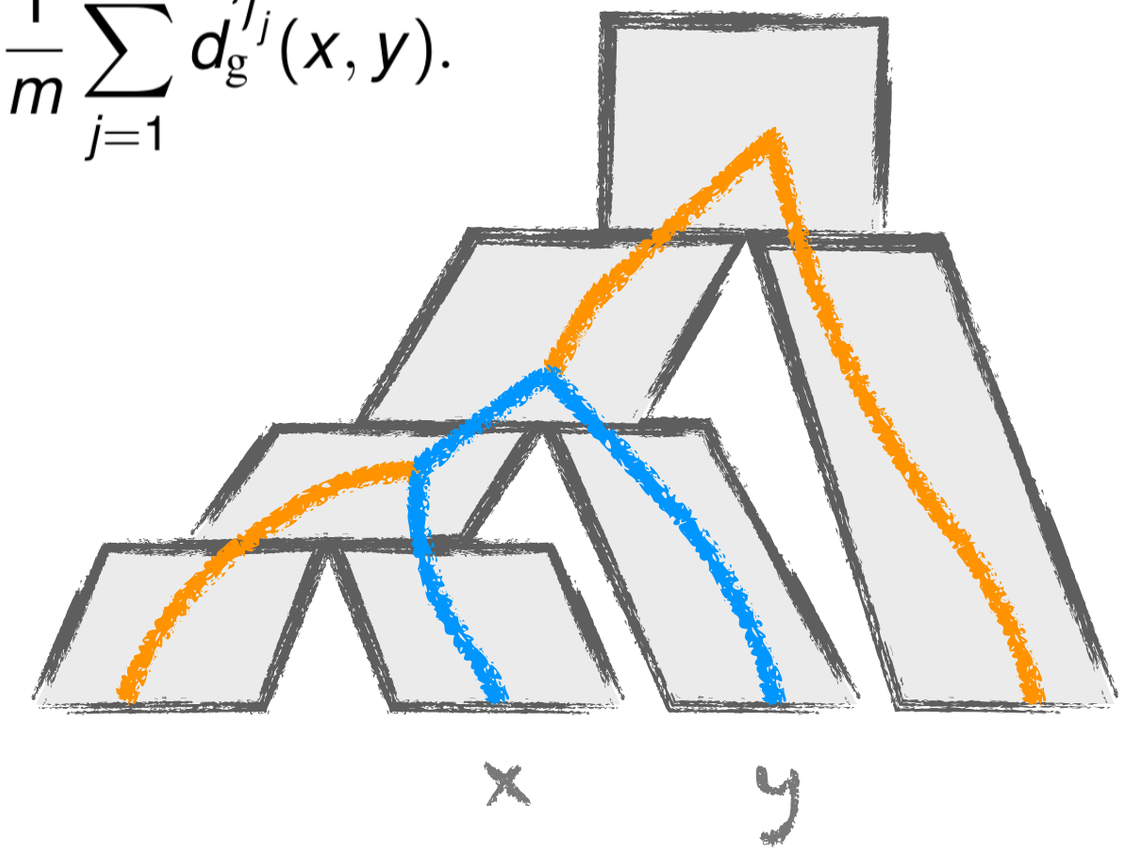
Internode distance

Definition

For any pair of species x, y and gene j, we let $d_{\mathrm{g}}^{\mathcal{T}_j}(x, y)$ be the *graph distance* between x and y on gene tree $\mathcal{T}_j$, i.e. the number of edges on the unique path between x and y. The *internode distance* between x and y is defined as the average graph distance across genes, i.e.

$$\hat{\delta}_{\mathrm{int}}^{m}(x, y) = \frac{1}{m} \sum_{j=1}^{m} d_{\mathrm{g}}^{\mathcal{T}_j}(x, y).$$

graph distance
between x and y = 3



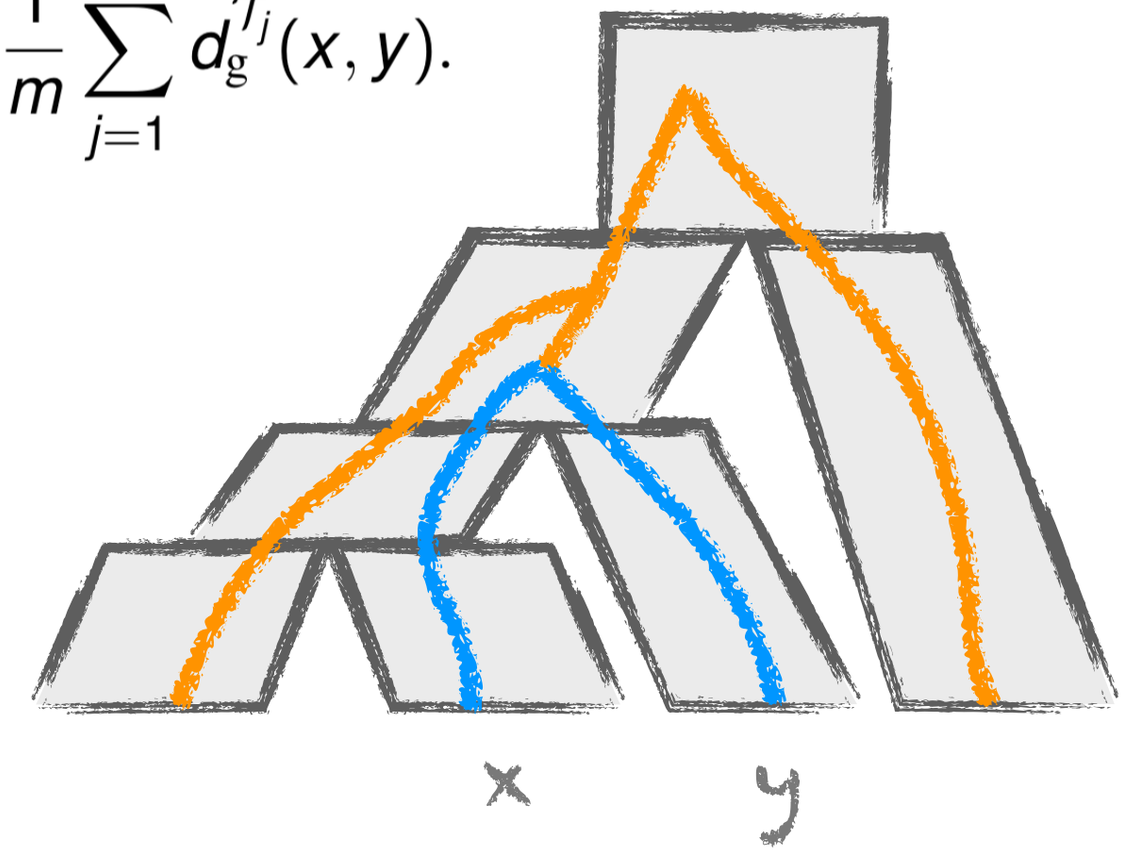
Internode distance

Definition

For any pair of species x, y and gene j, we let $d_{\mathrm{g}}^{\mathcal{T}_j}(x, y)$ be the *graph distance* between x and y on gene tree $\mathcal{T}_j$, i.e. the number of edges on the unique path between x and y. The *internode distance* between x and y is defined as the average graph distance across genes, i.e.

$$\hat{\delta}_{\mathrm{int}}^{m}(x, y) = \frac{1}{m} \sum_{j=1}^{m} d_{\mathrm{g}}^{\mathcal{T}_j}(x, y).$$

graph distance between x and y = 2

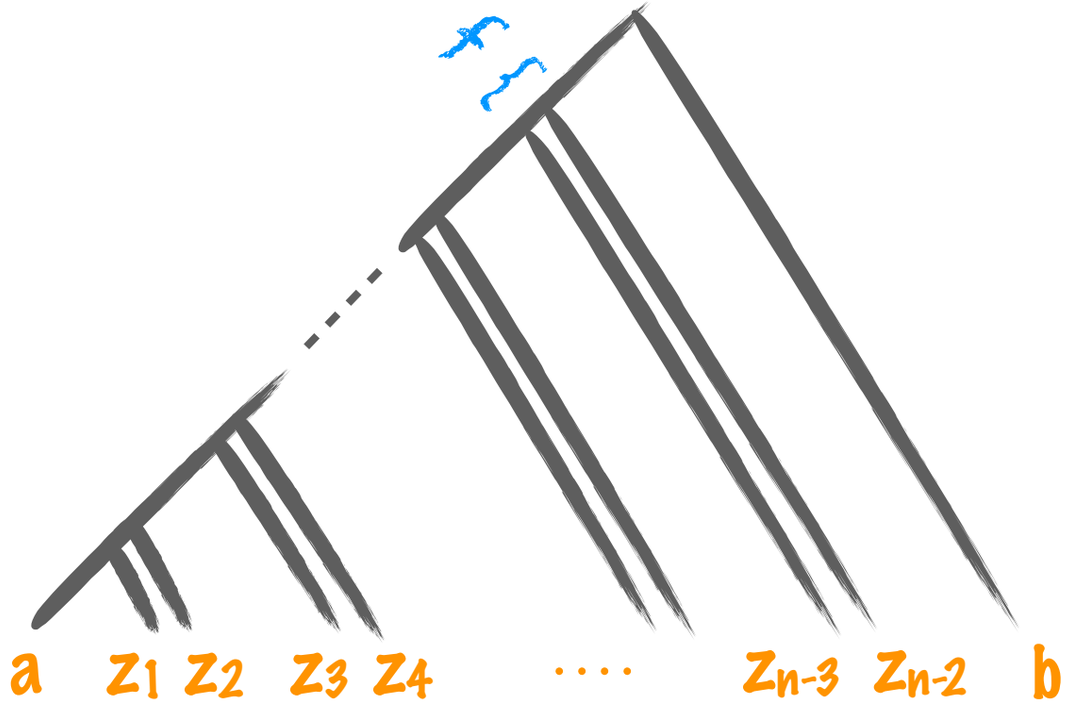


Variance of internode

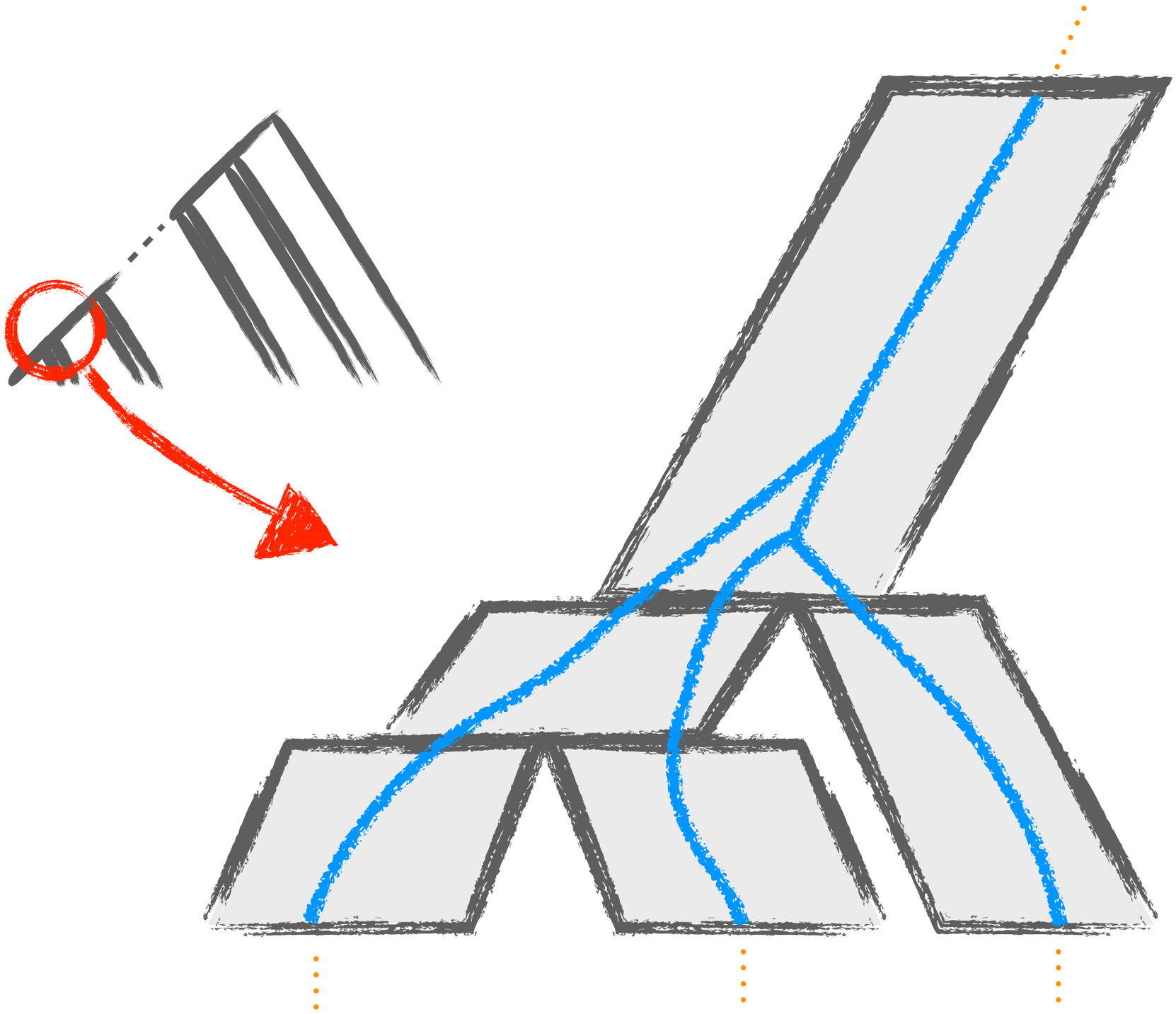
$\hat{\delta}_{\text{int}}^{(m)}(a,b)$ = mean graph distance from a to b over m genes

Theorem (R. (2018))

There is a species tree with n leaves, shortest branch length f, and a pair of species a and b such that $\text{Var}[\hat{\delta}_{\text{int}}^{(m)}(a,b)] = \Omega(\frac{n}{m})$.



Proof by picture



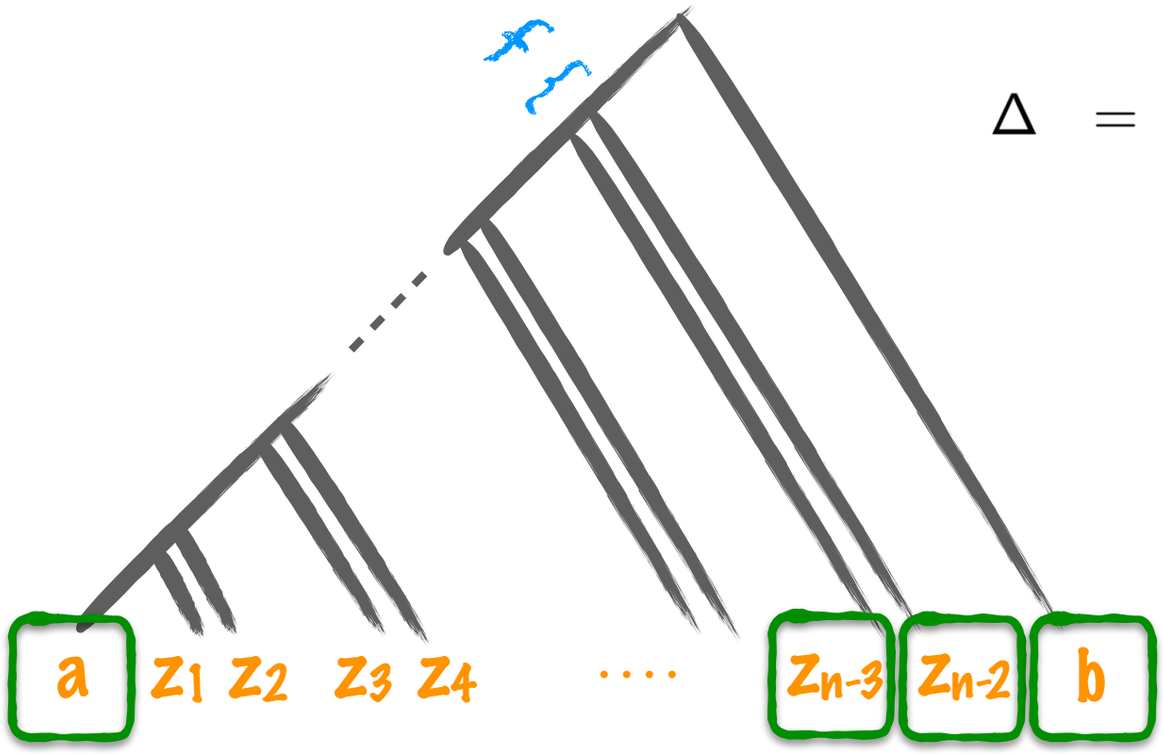
The impact of correlation?

$\hat{\delta}^{(m)}_{\text{int}}(a,b) =$ mean graph distance from a to b over m genes

Theorem (R. (2018))

There is a species tree with n leaves, shortest branch length f, and a pair of species a and b such that $\mathrm{Var}[\hat{\delta}^{(m)}_{\text{int}}(a,b)] = \Omega(\frac{n}{m})$.

However, on the same example, the variance of the "four-point formula" is significantly smaller because of correlation.



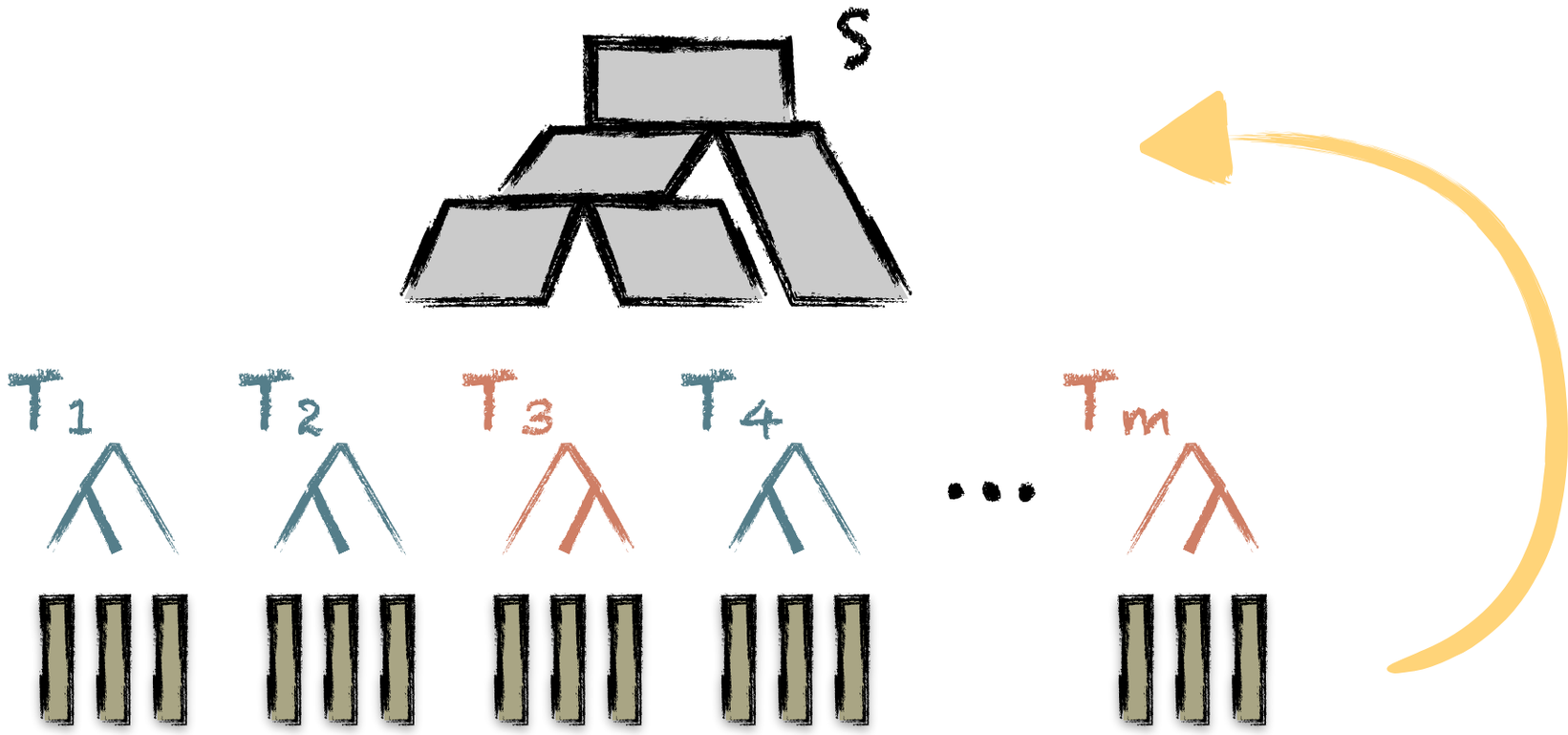
$$\begin{aligned}\Delta \quad = \quad & \hat{\delta}^{(m)}_{\text{int}}(a,b) + \hat{\delta}^{(m)}_{\text{int}}(z_{n-3},z_{n-2}) \\ & -\hat{\delta}^{(m)}_{\text{int}}(a,z_{n-2}) - \hat{\delta}^{(m)}_{\text{int}}(z_{n-3},b)\end{aligned}$$

$$\mathrm{E}[\Delta] = \Theta(f)$$

$$\mathrm{Var}[\Delta] = \Theta\left(\frac{1}{m}\right)$$

The full model: MSC-JC

- Species tree: S
- For each gene g (independently and identically),
  - Generate a gene tree T_g for g using the multispecies coalescent on S
  - Generate sequence data of length k on T_g using a substitution model
- Goal: reconstruct S from the sequences



MLE on concatenation is not consistent

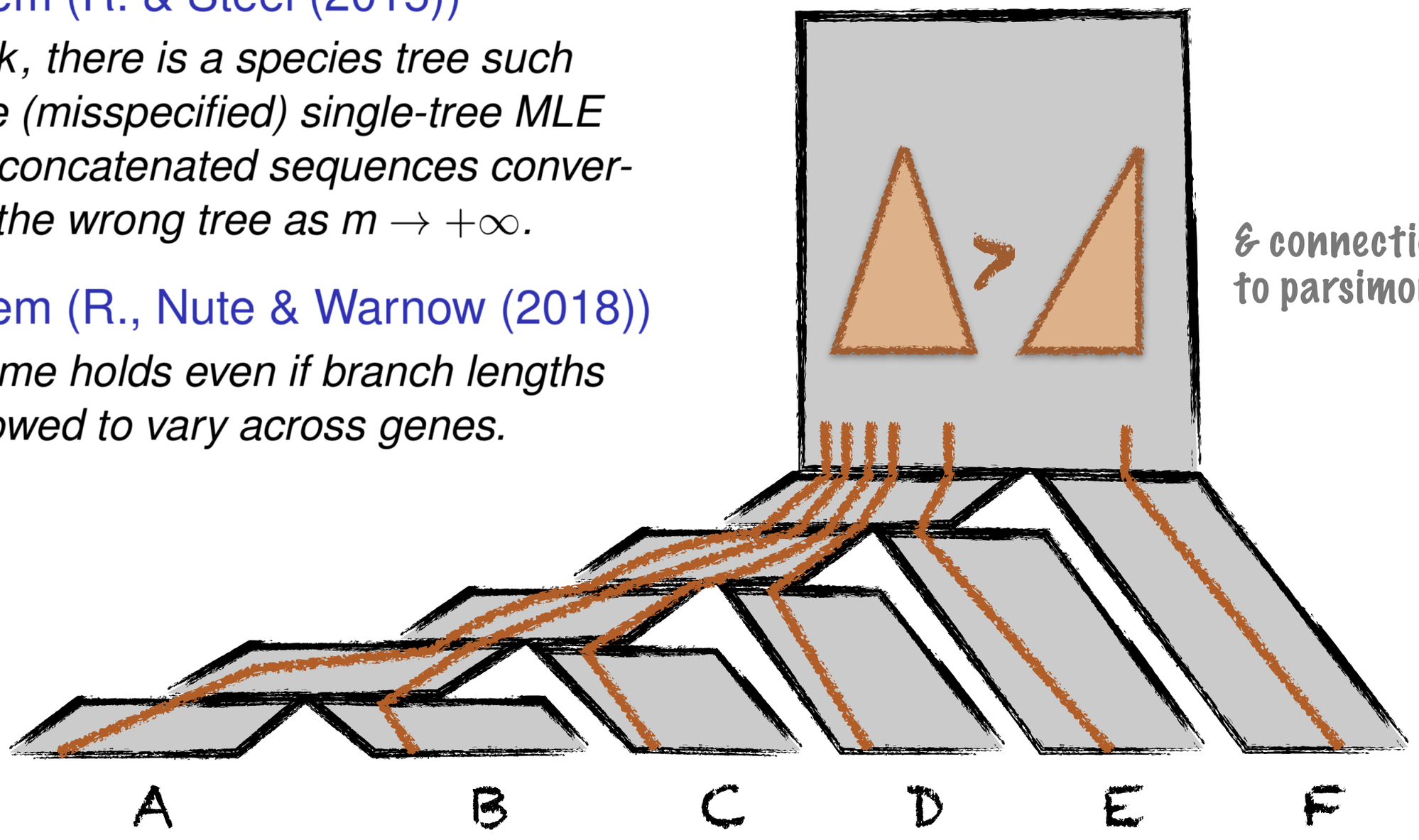
m: number of genes; *k*: gene length

Theorem (R. & Steel (2015))

For all k, there is a species tree such that the (misspecified) single-tree MLE on the concatenated sequences converges to the wrong tree as m $\to +\infty$.

Theorem (R., Nute & Warnow (2018))

The same holds even if branch lengths are allowed to vary across genes.



& connection to parsimony

Concatenation revisited

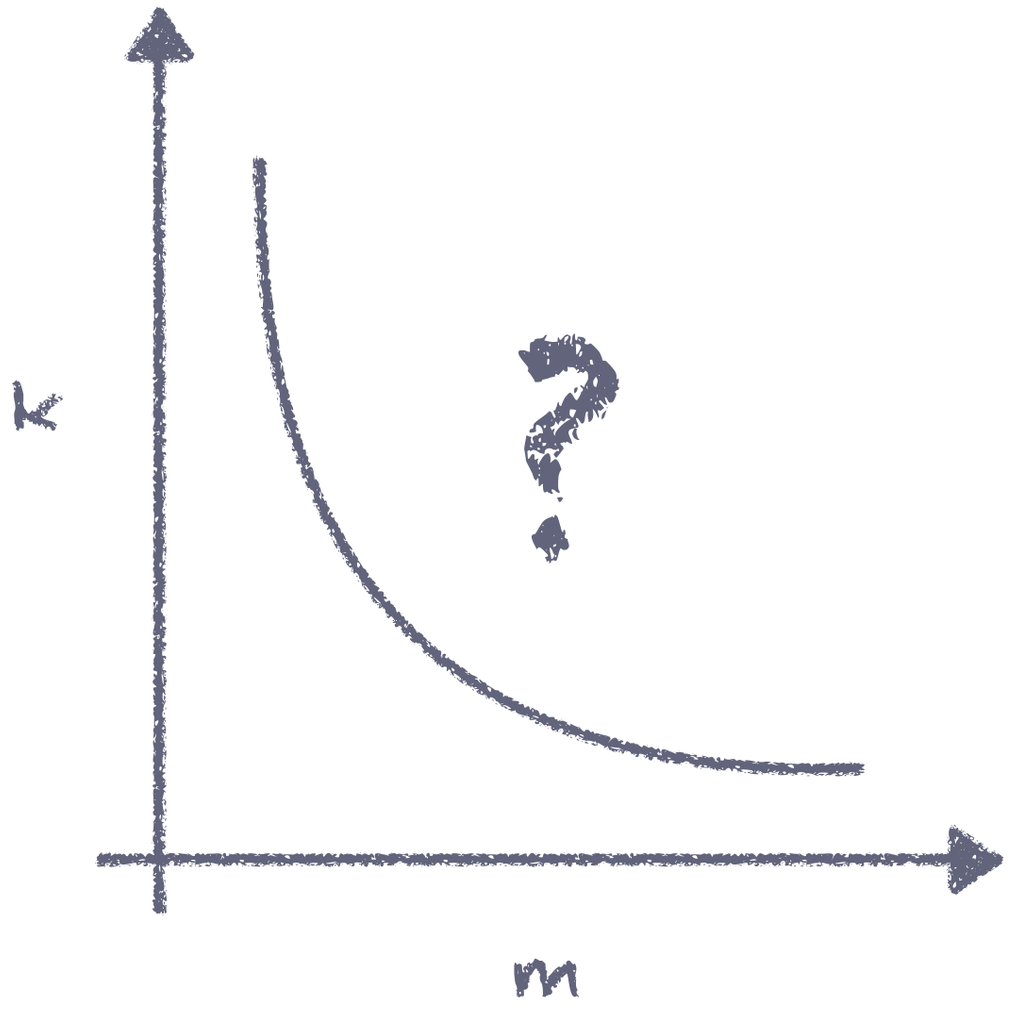
Theorem (Dasarathy, Nowak & R. (2015))

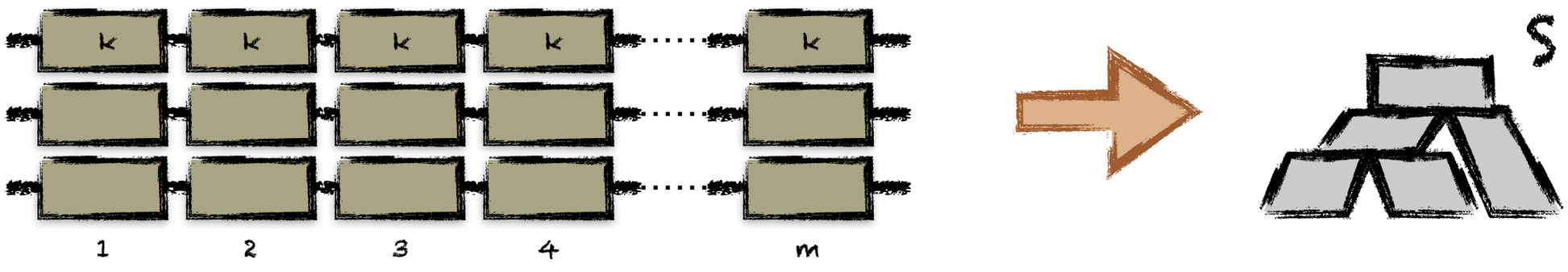
Under MSC-JC with varying population sizes and lineage-specific mutation rates, the species tree is identifiable from the distribution of a single site.

The proof is based on showing that a Jukes-Cantor version of the log-det distance **over the concatenated sequences** is a tree metric for the species tree. It also gives a consistent reconstruction method **for any gene length**.

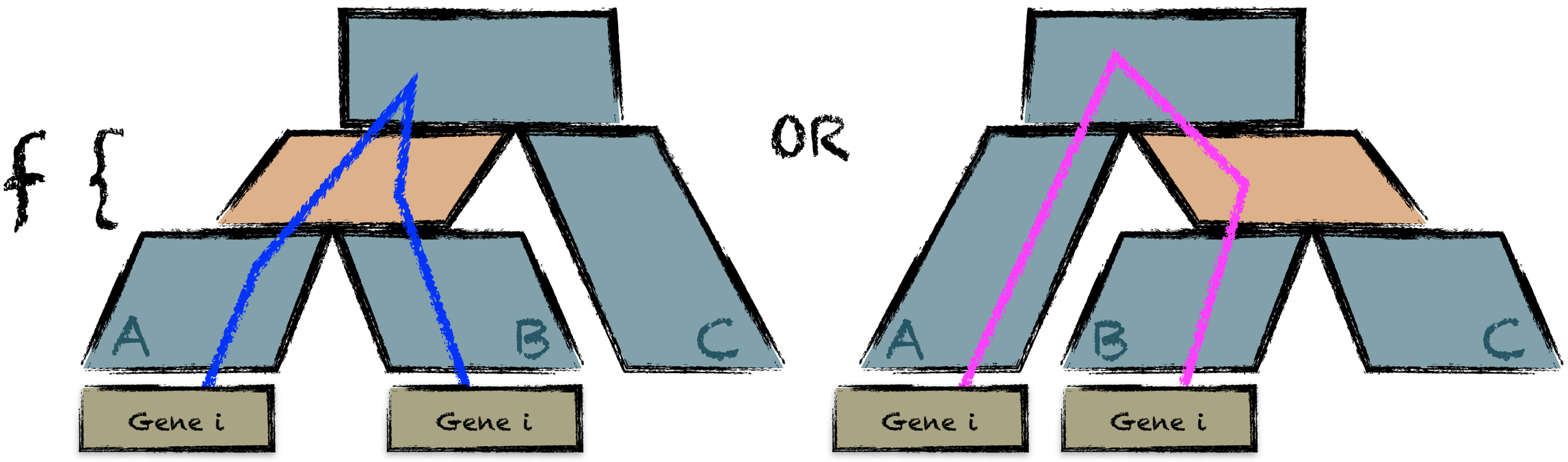
Identifiability results under other assumptions have also been obtained (Chifman & Kubatko (2014); Long & Kubatko (2017); Allman et al. (2018)).

How much data is needed?





Information-theoretic lower bound on the data requirement

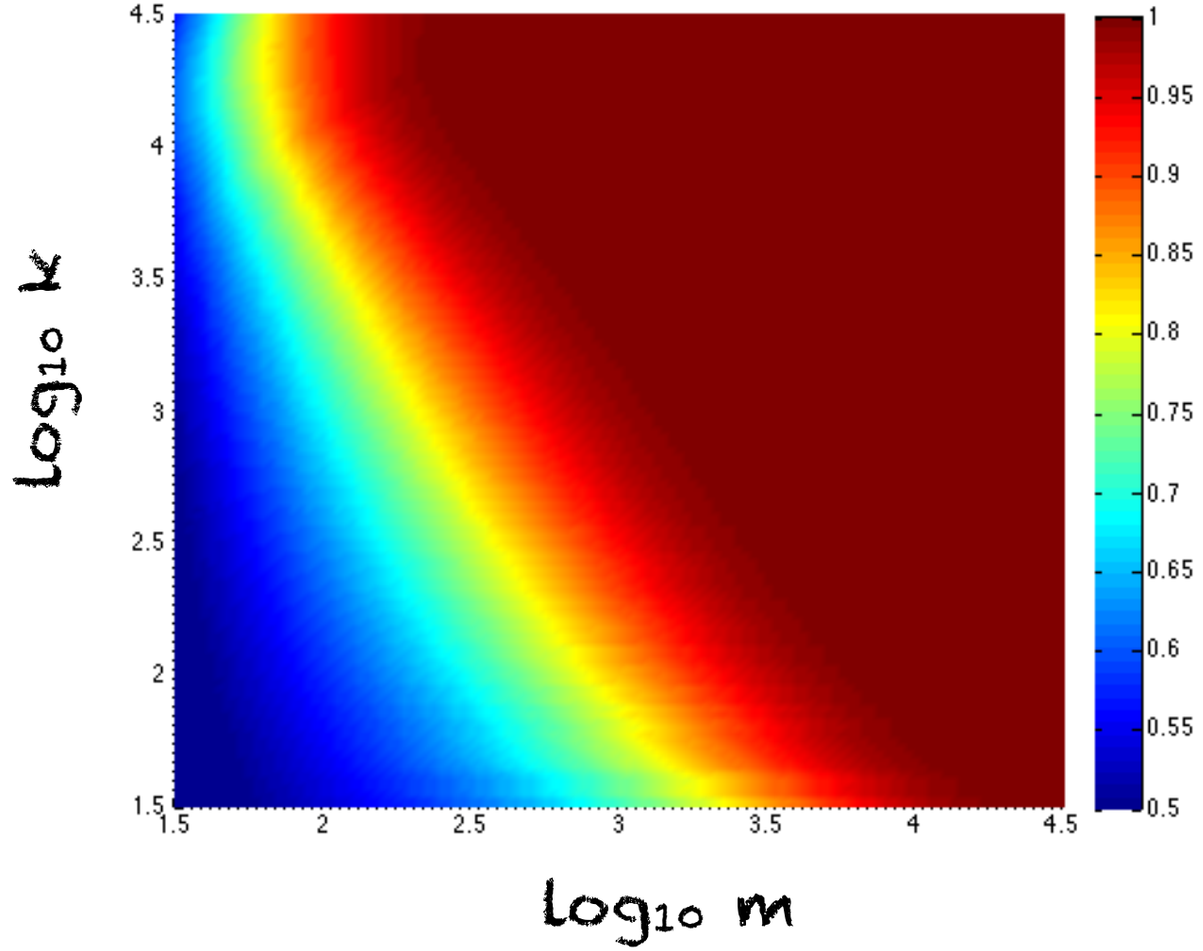


An unexpected trade-off

m: number of genes; k: gene length; f: shortest branch

Theorem (Mossel & R. (2015, 2018))

Under MSC-JC, reconstruction with high probability requires $m\sqrt{k} \geq C_0 f^{-2}$ *when* $k \leq C_1 f^{-2}$. *(Achieved "under some conditions" (Dasarathy, Nowak, Mossel & R. (2018)).)*

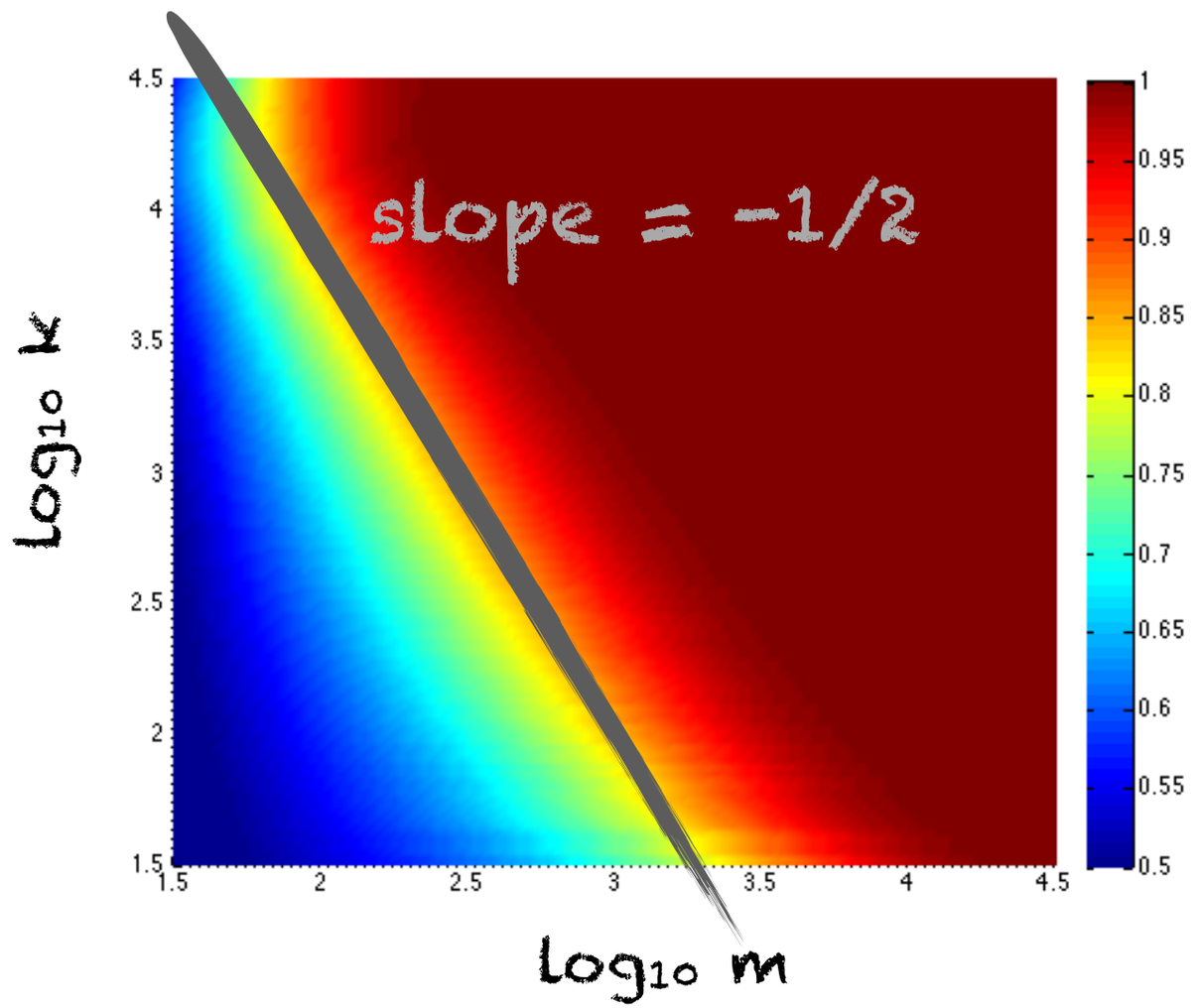


An unexpected trade-off

m: number of genes; k: gene length; f: shortest branch

Theorem (Mossel & R. (2015, 2018))

Under MSC-JC, reconstruction with high probability requires $m\sqrt{k} \geq C_0 f^{-2}$ *when* $k \leq C_1 f^{-2}$. *(Achieved "under some conditions" (Dasarathy, Nowak, Mossel & R. (2018)).)*



Another source of discordance: horizontal gene transfer (HGT)

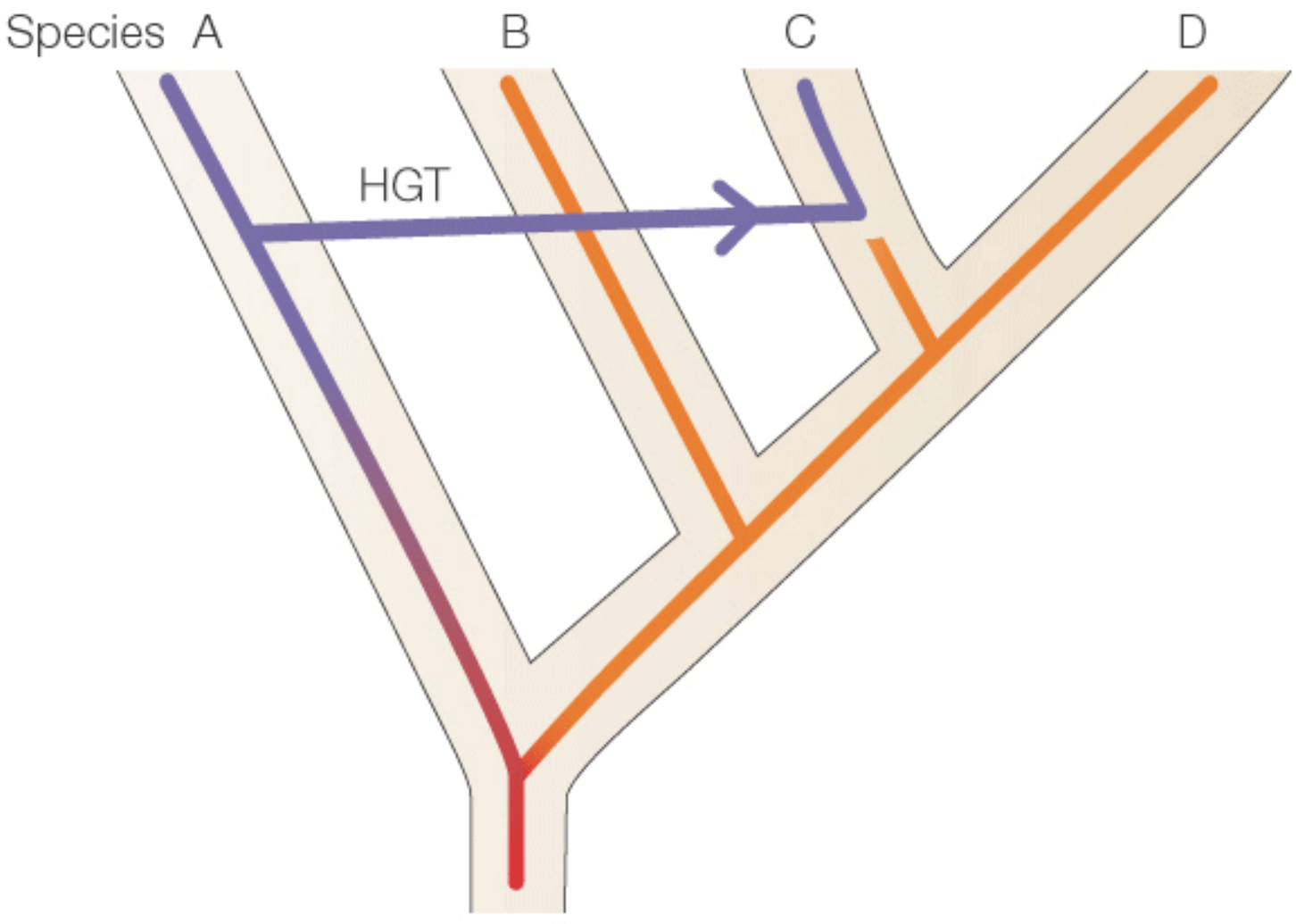


Figure from Gogarten & Townsend Nature Reviews Microbiology 3, 679-687 (2005)

A stochastic model of HGT

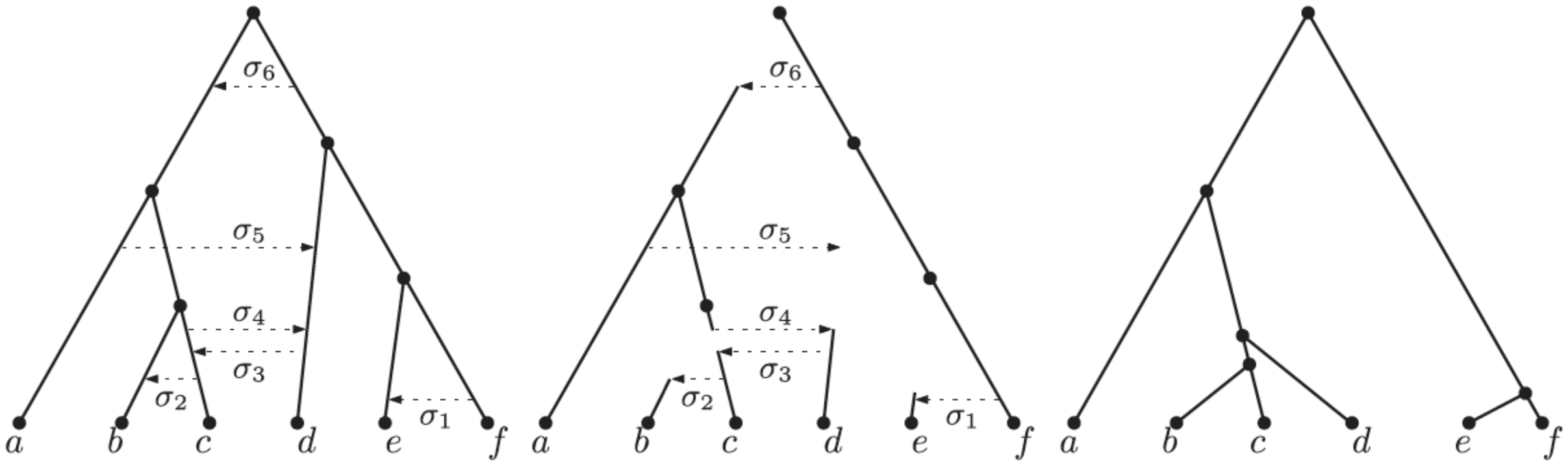
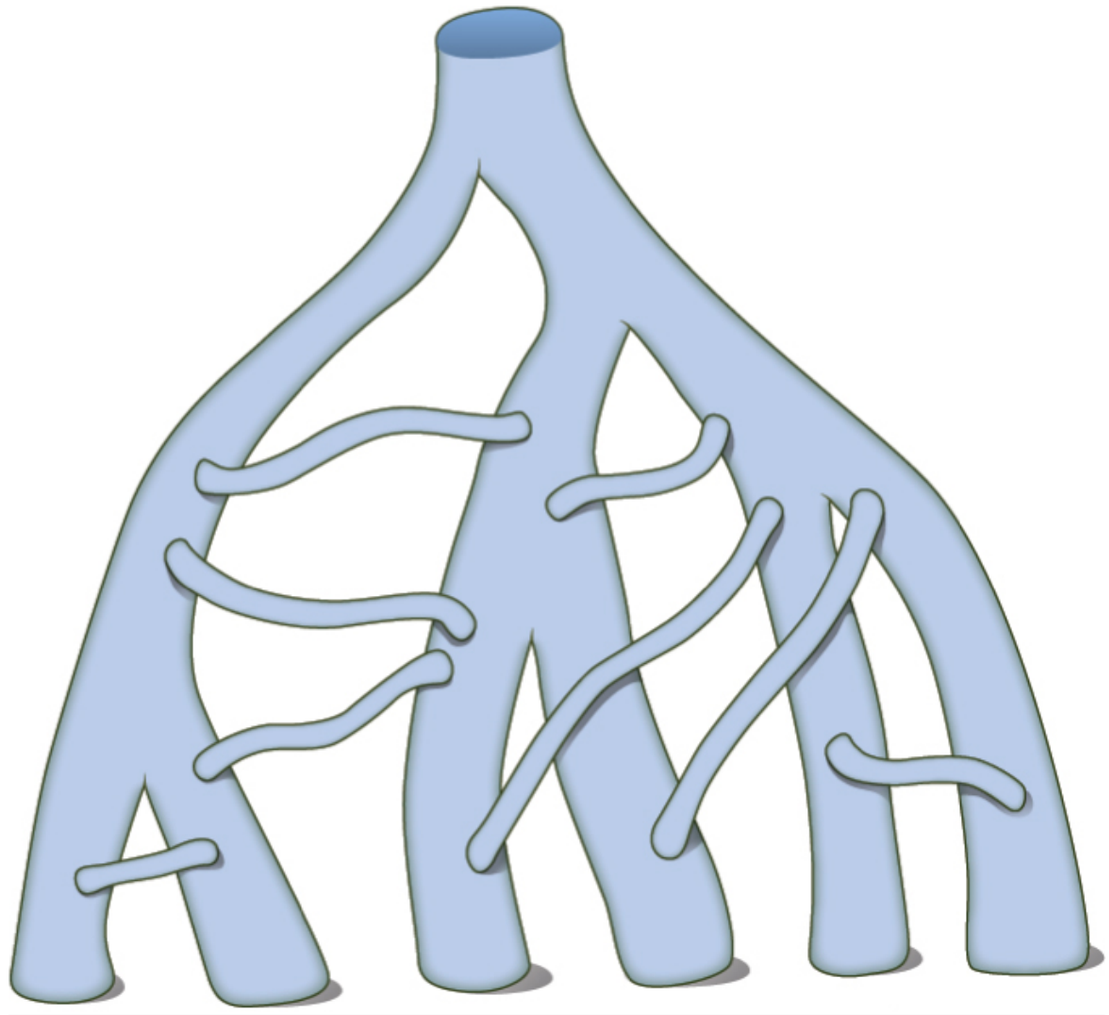


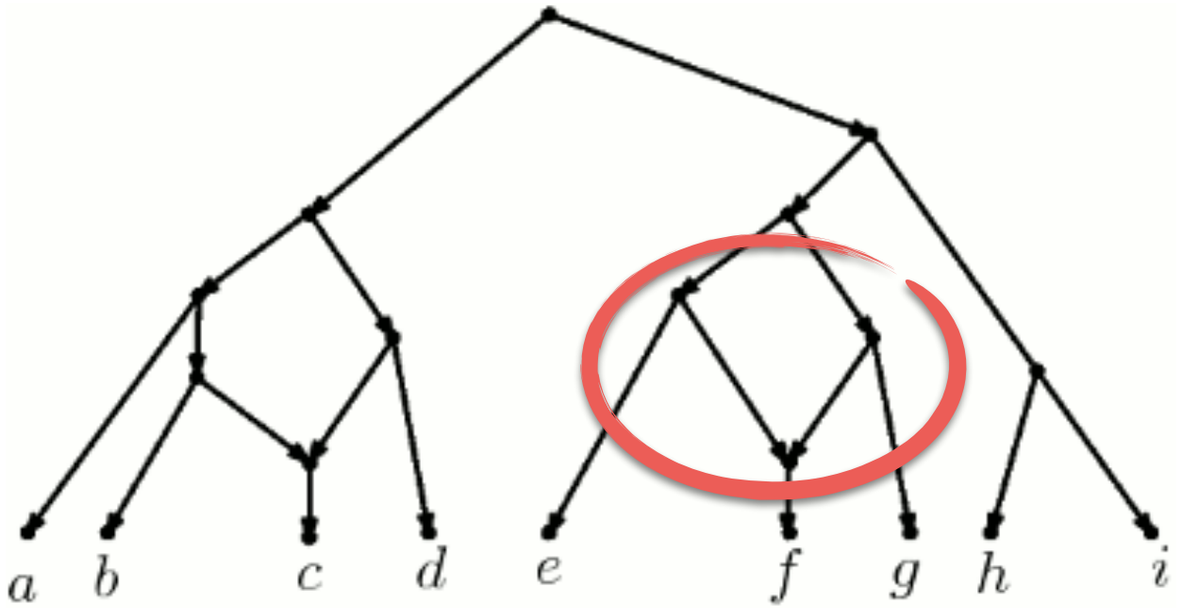
Figure from Steel et al., JTB (2013)



Is the Tree of Life even a tree?

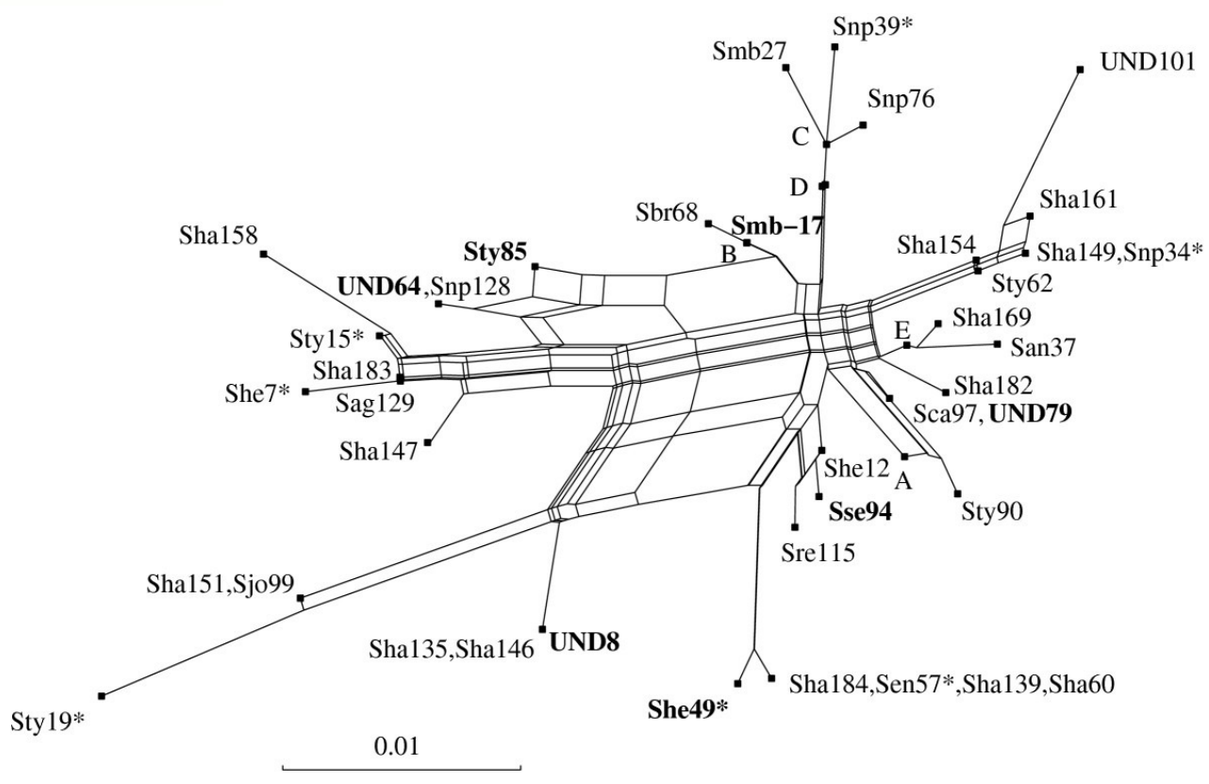
E.g., hybridization?

Beyond trees



Phylogenetic network

Split network

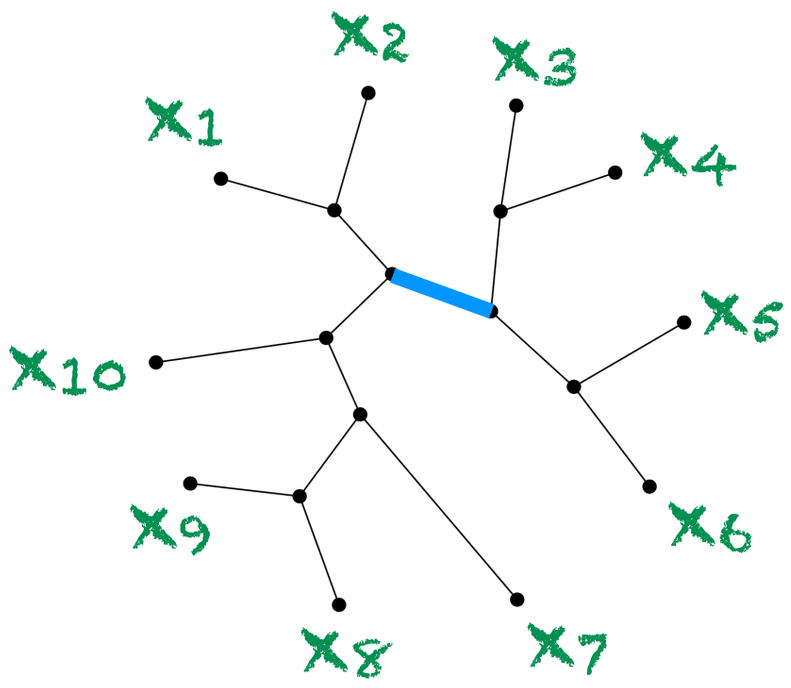


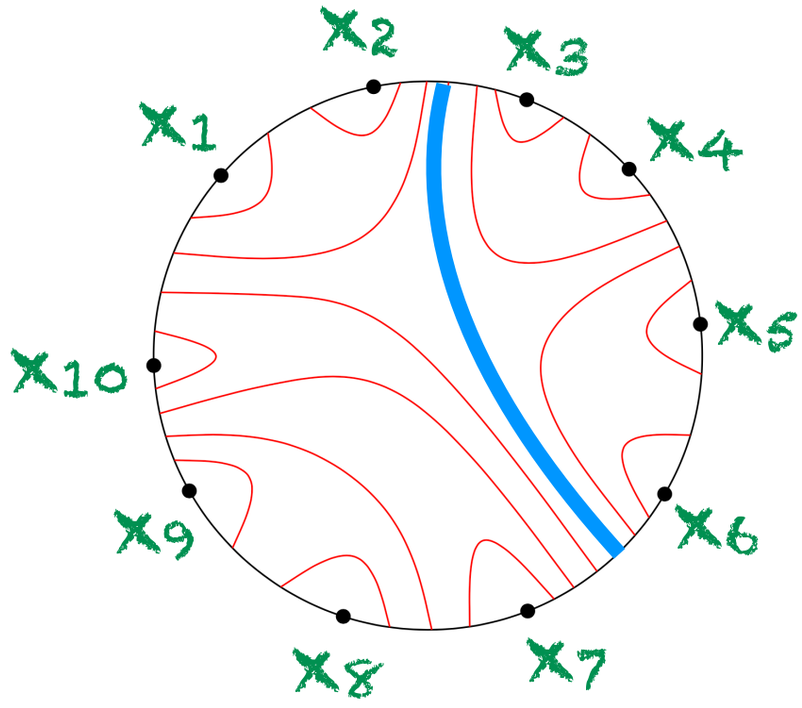
Trees as circular split systems

Definition

A collection of X-splits $\mathcal{S}$ is called *circular* if there exists a linear ordering $(x_1, \ldots, x_n)$ of the elements of X such that each split $S \in \mathcal{S}$ has the form: for some $1 < p \leq q \leq n$,

$$S = \{ \{x_p, \ldots, x_q\}, X - \{x_p, \ldots, x_q\} \}.$$

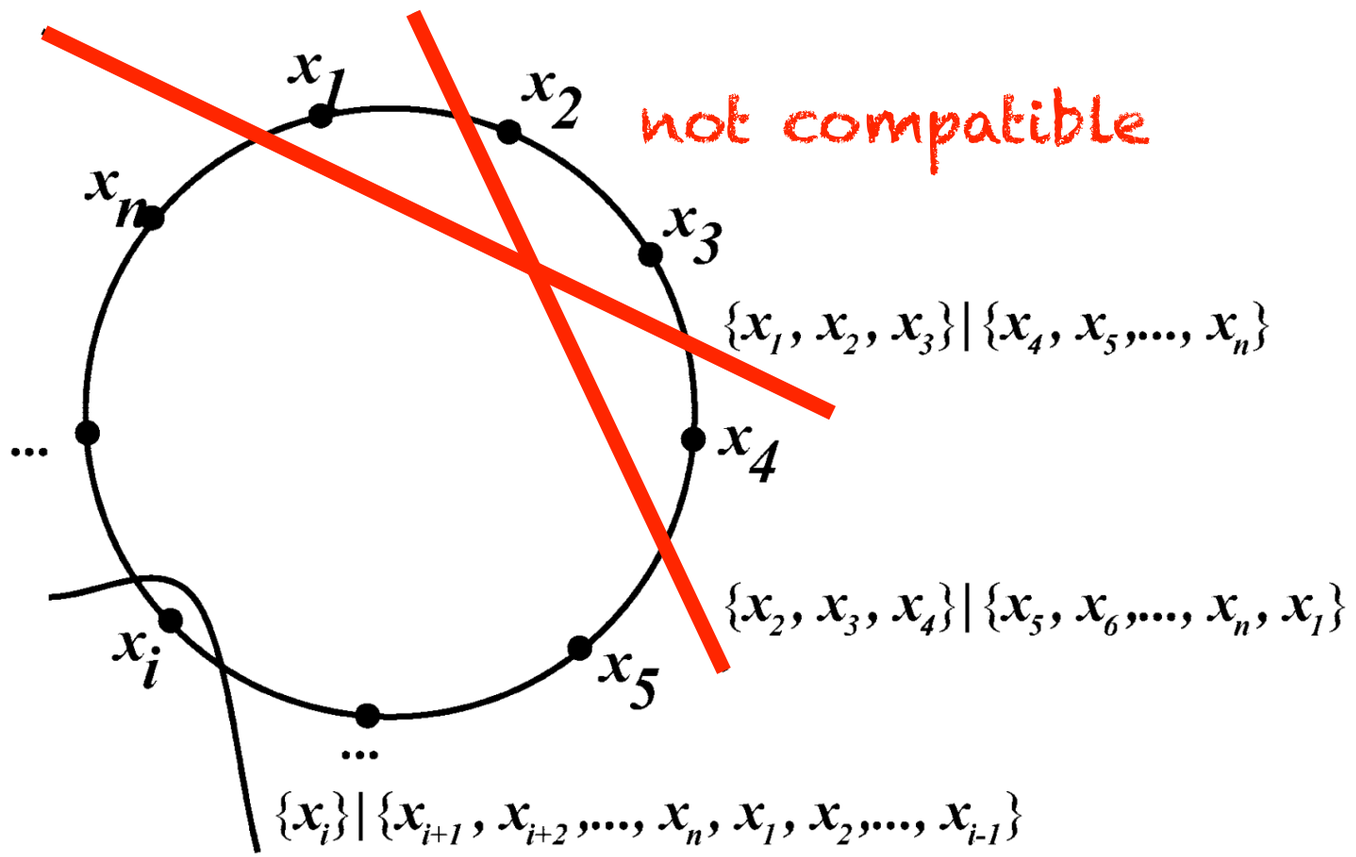




General circular split systems

Definition

We say that $\mathcal{N} = (X, \mathcal{S}, w)$ is a *split network* on a set X if $\mathcal{S}$ is a set of splits on X and $w : \mathcal{S} \to (0, \infty)$ is a weight function. A split network $\mathcal{N} = \{X, \mathcal{S}, w\}$ is *circular network* if $\mathcal{S}$ is.

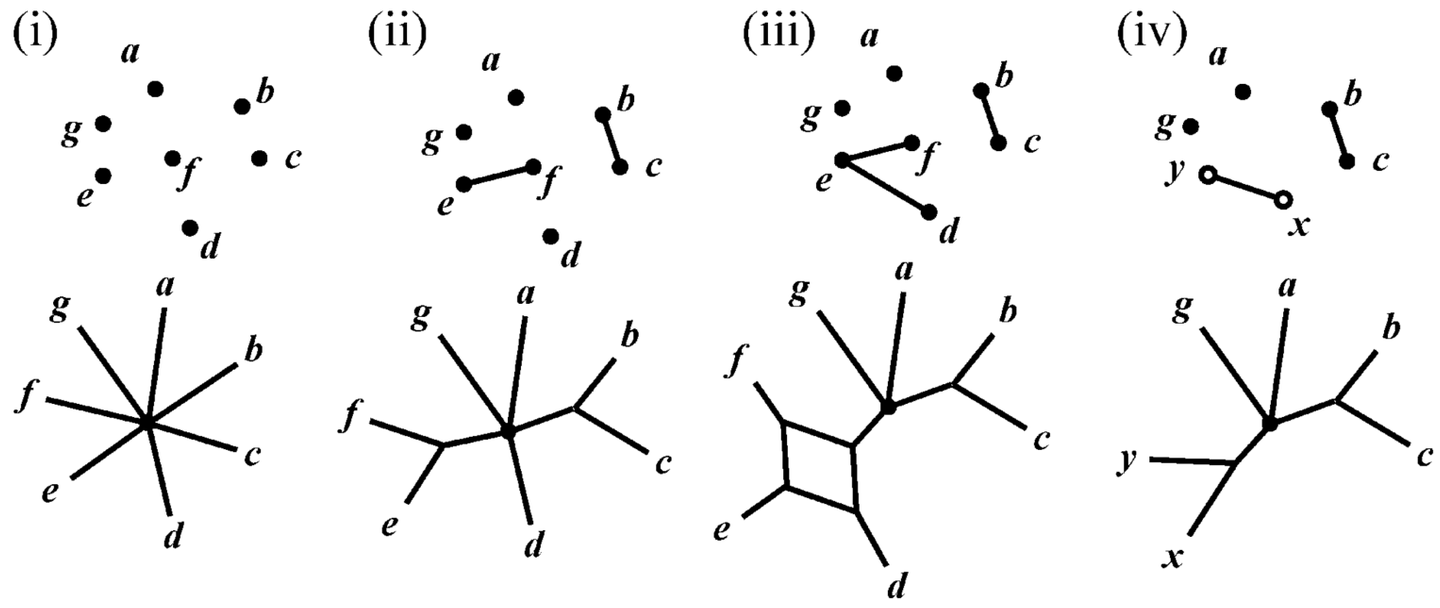


Why circular networks I: Neighbor-Net [Bryant-Moulton'04]

0	3	8	1	10	8
	0	9	2	11	9
		0	7	8	6
			0	9	7
				0	4
					0

$$Q(C_i, C_j) = (m-2)d(C_i, C_j) - \sum_{\substack{k=1 \\ k \neq i}}^{m} d(C_i, C_k) - \sum_{\substack{k=1 \\ k \neq j}}^{m} d(C_j, C_k).$$

where $d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} d_{xy}.$



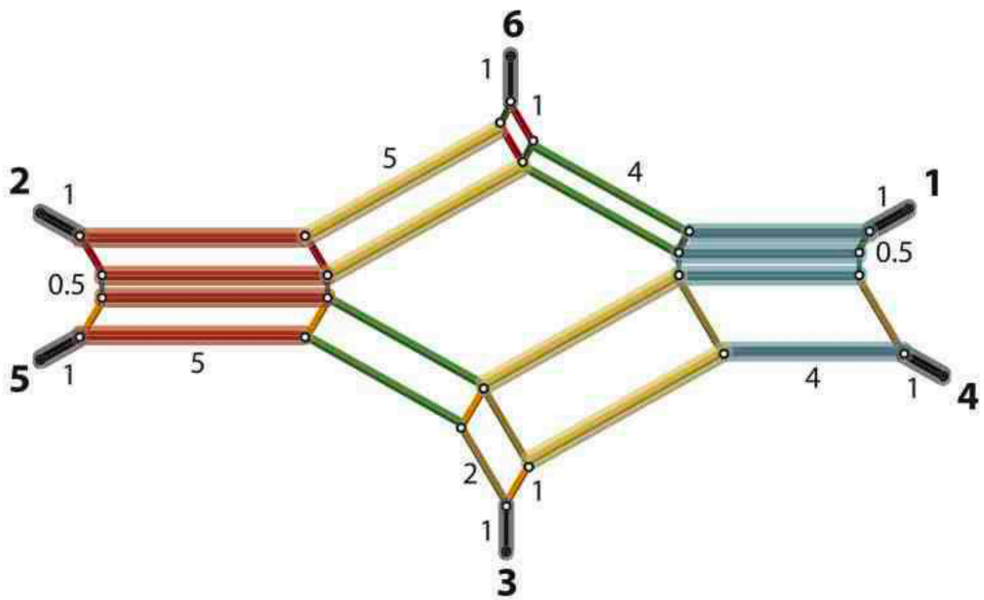
Split metrics

Definition

Let $\mathcal{N} = (X, \mathcal{S}, w)$ be a split network. The dissimilarity $\delta : X \times X \to [0, \infty)$ defined as follows

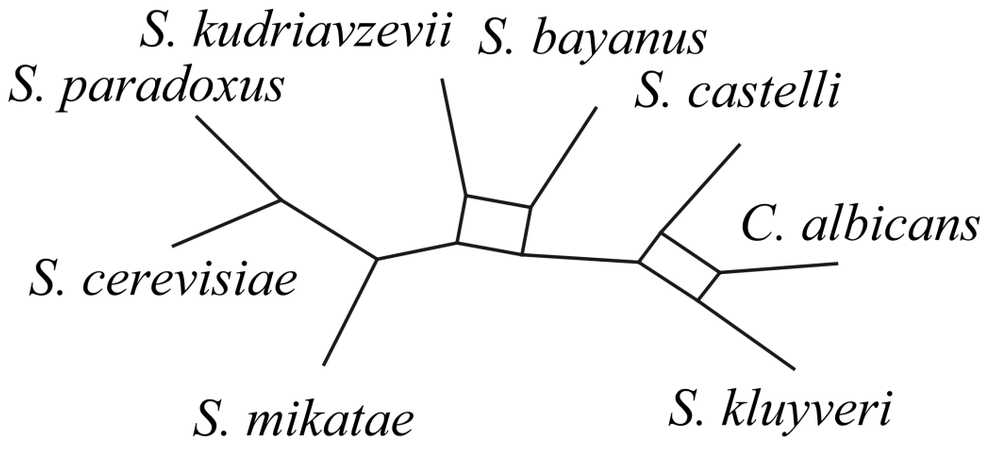
$$\delta(x, y) = \sum_{S \in \mathcal{S}|_{x,y}} w(S),$$

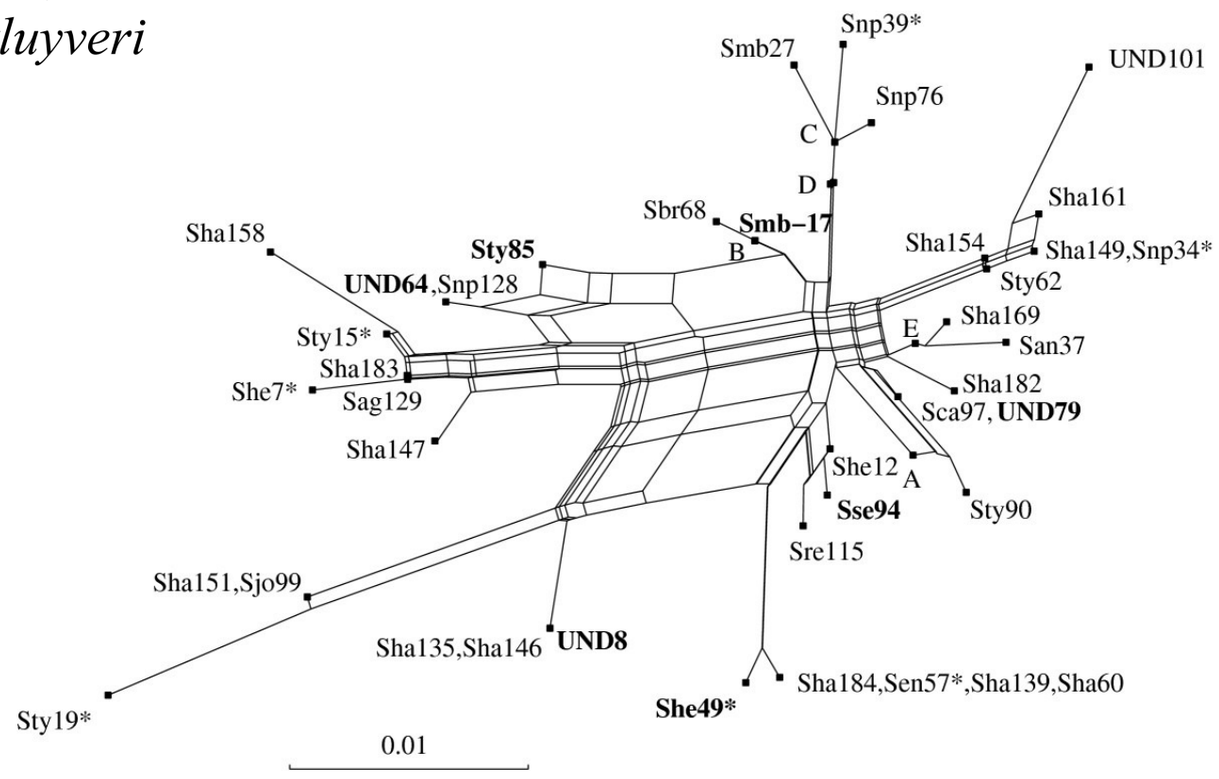
for all $x, y \in X$, is referred to as the metric associated to $\mathcal{N}$, where $\mathcal{S}|_{x,y}$ is the collection of splits in $\mathcal{S}$ "separating" x and y.



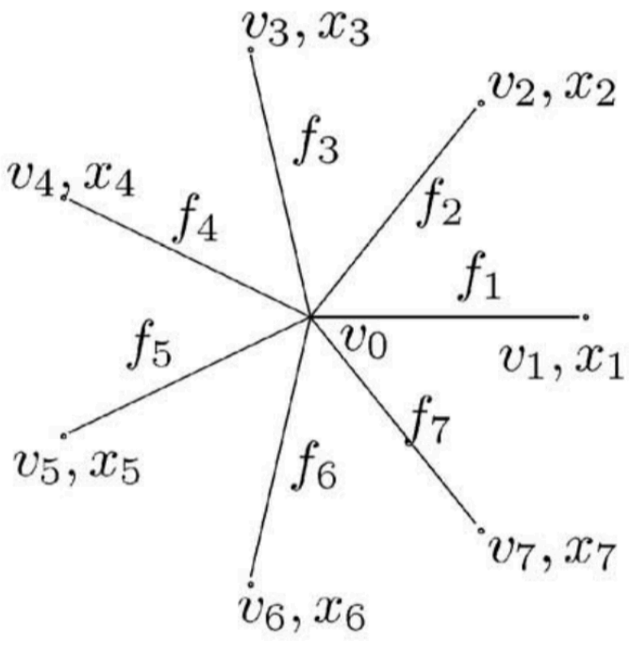
	1	2	3	4	5	6
1	0	21.5	15	5	22	11
2	21.5	0	15.5	23.5	4.5	12.5
3	15	15.5	0	12	13	16
4	5	23.5	12	0	23	14
5	22	4.5	13	23	0	15
6	11	12.5	16	14	15	0

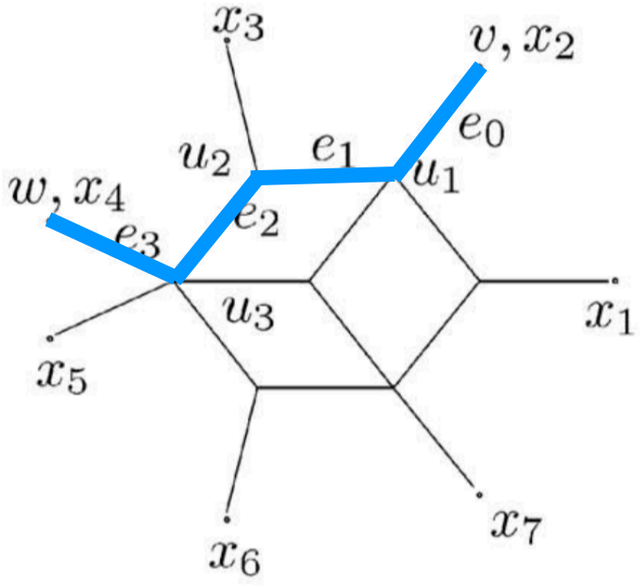
Why circular networks II: Outer-labeled planar splits graph [Wetzel'95, Dress-Huson'04]

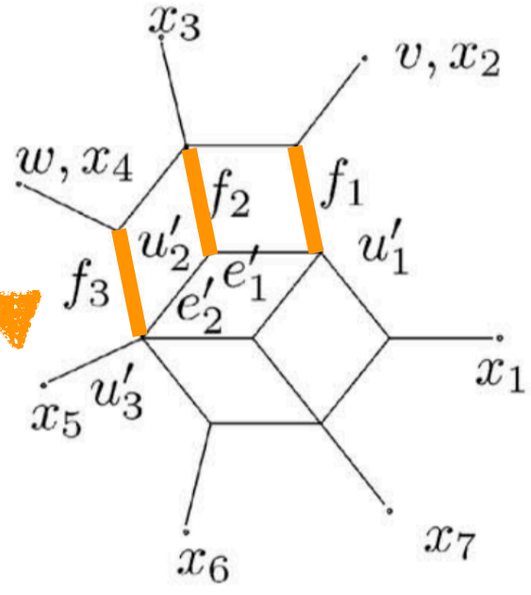




Why circular networks II: Outer-labeled planar splits graph [Wetzel'95, Dress-Huson'04]





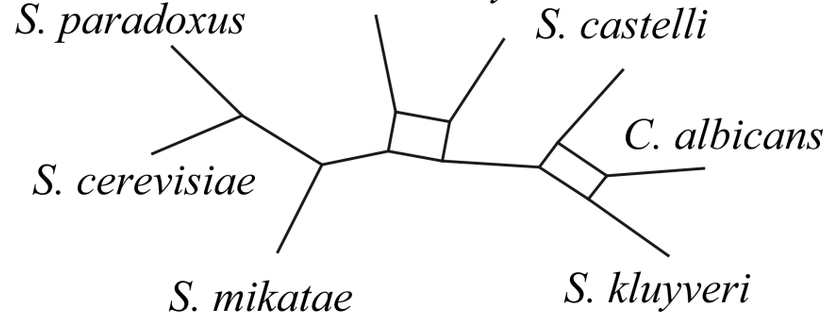


$$\frac{\{x_2, x_3, x_4\}}{\{x_1, x_5, x_6, x_7\}}$$

A more robust algorithm: overview

"Theorem" [Roch-Wang'18]: We give an efficient reconstruction algorithm for circular networks from distorted metrics with a much smaller radius of accuracy than previous methods.

0	3	8	1	10	8
	0	9	2	11	9
		0	7	8	6
			0	9	7
				0	4
					0



Jukes-Cantor formula

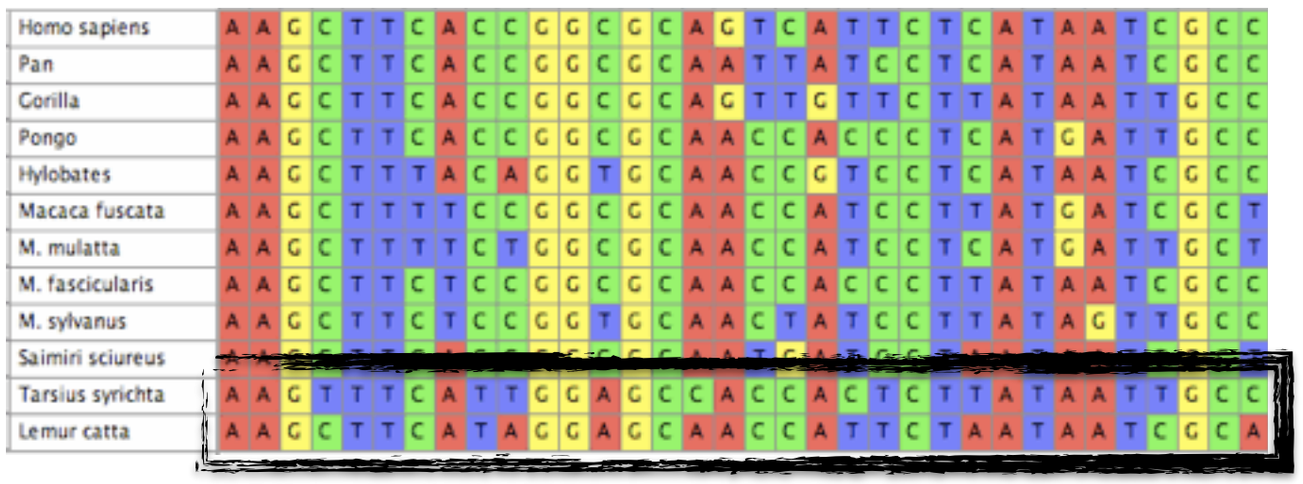
Definition

Let H^n_{xy} be the Hamming distance between the sequences (of length n) at the leaves x and y (i.e., the number of changes). The *Jukes-Cantor distance formula* is

$$\hat{\delta}_n(x,y) = -\frac{3}{4}\log\left(1 - \frac{4}{3}\cdot\frac{H^n_{xy}}{n}\right).$$

Theorem

As $n \to \infty$, $\hat{\delta}_n(x,y)$ converges a.s. to $\mu \cdot t_{xy}$, where μ is the mutation rate and t_{xy} is the "time elapsed between x and y."

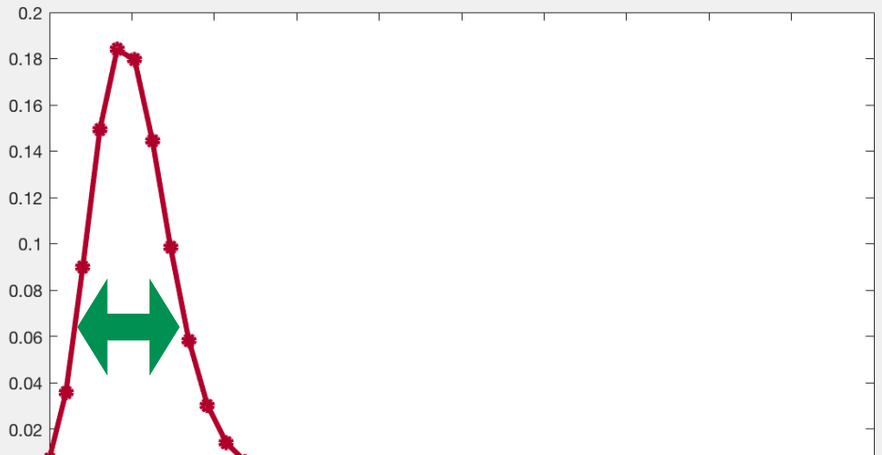


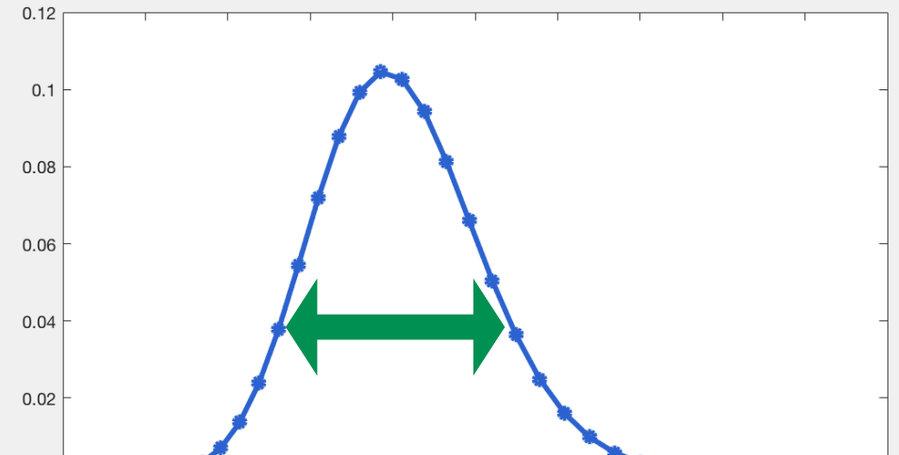
Variance of Jukes-Cantor formula increases with evolutionary distance

t = 50

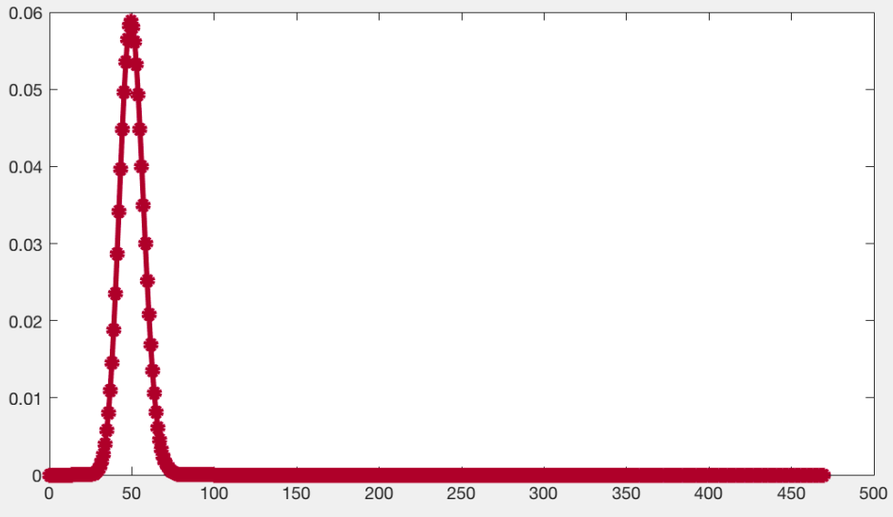
t = 200

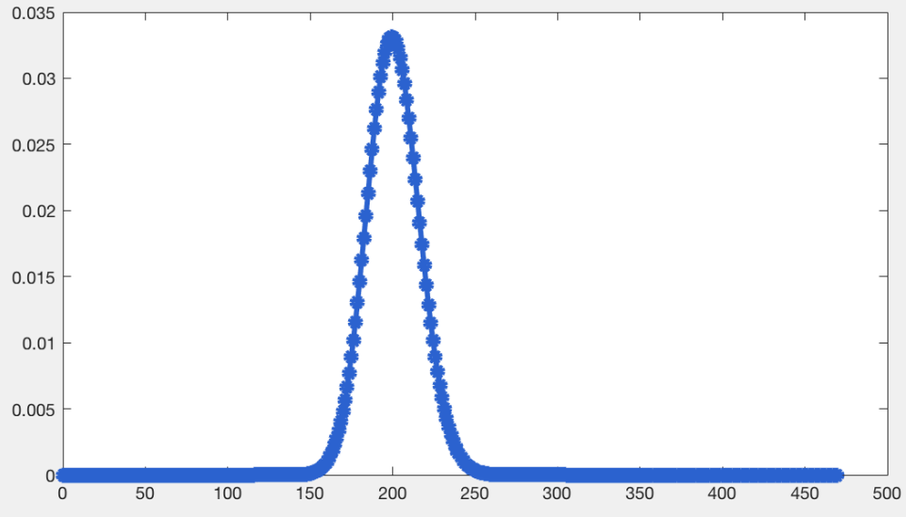
n = 100





n = 1000





Distorted metrics

Definition

Suppose $\mathcal{N} = (\mathcal{X}, \mathcal{S}, w)$ is a split network with associated metric δ. We say that $\hat{\delta} : X \times X \to [0, +\infty]$ is a (τ, R)-*distorted metric* of $\mathcal{N}$ if $\hat{\delta}$ is "accurate" on "short" distances, i.e. $\forall x, y \in X$

$$\delta(x, y) \wedge \hat{\delta}(x, y) < R + \tau \implies |\delta(x, y) - \hat{\delta}(x, y)| < \tau.$$

We refer to τ and R as the *tolerance* and *accuracy radius*.

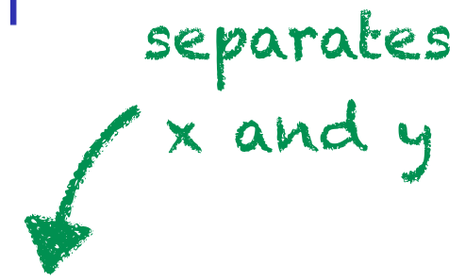
Tree case

Definition

The *mininum weight* of $\mathcal{N}$ is $\epsilon_{\mathcal{N}} = \min\{w(S) : S \in \mathcal{S}\}$.

Definition (R.-Wang (2018); generalization of notion introduced in Erdös-Steel-Székely-Warnow (1999))

The *chord depth* of a split $S \in \mathcal{S}$ is



$$\Delta_{\mathcal{N}}(S) = \min\left\{\delta(x, y; \mathscr{C}_{\mathcal{N}}(S)) : x, y \in X \text{ such that } S \in \mathcal{S}|_{x,y}\right\},$$

where we restricted $\delta(x, y)$ to the splits compatible to S. The *chord depth* of $\mathcal{N}$ is $\Delta_{\mathcal{N}} = \max\{\Delta_{\mathcal{N}}(S) : S \in \mathcal{S}\}$.

Theorem (Daskalakis-Mossel-R. (2011); implicit in Erdös-Steel-Székely-Warnow (1999))

If $\mathcal{N} = (\mathcal{X}, \mathcal{S}, w)$ is compatible, then a (τ, R)-distorted metric with $\tau < \frac{1}{4}\epsilon_{\mathcal{N}}$ and $R > 2\Delta_{\mathcal{N}} + \frac{5}{4}\epsilon_{\mathcal{N}}$ can be used to reconstruct $\mathcal{N}$ in polynomial time.

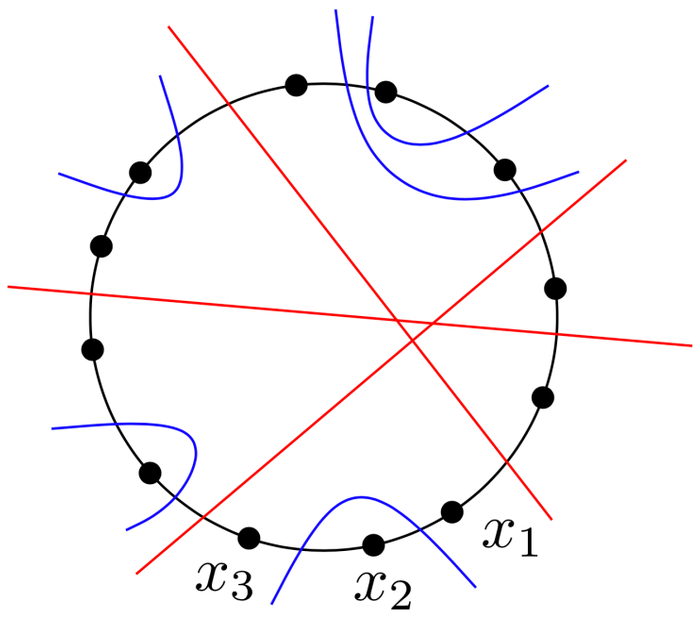
Incompatibility

Definition (R.-Wang (2018))

The *incompatible weight* of a split $S \in \mathcal{S}$ is

$$\Omega_{\mathcal{N}}(S) = \sum_{S' \in \mathscr{I}(S)} w(S'),$$

where the sum is over splits incompatible with S. The *maximum incompatibility* of $\mathcal{N}$ is $\Omega_{\mathcal{N}} = \max\{\Omega_{\mathcal{N}}(S) : S \in \mathcal{S}\}$.



Main results

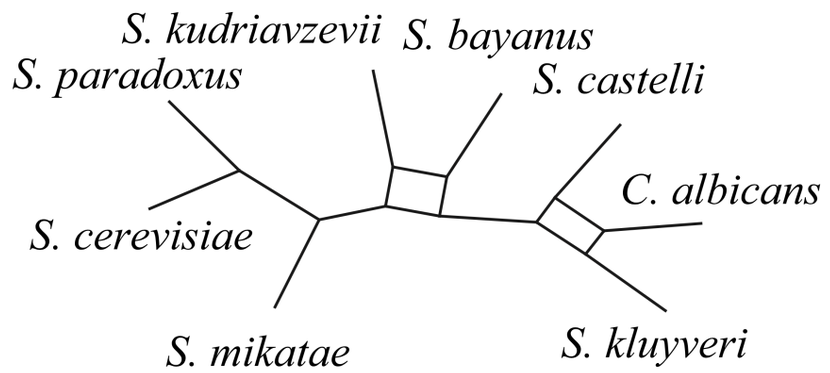
Theorem (R.-Wang'18)

Suppose $\mathcal{N} = (\mathcal{X}, \mathcal{S}, w)$ *is a circular network. Given a* (τ, R)*-distorted metric for* $\mathcal{N}$ *with*

$$\tau < \frac{1}{4}\epsilon_{\mathcal{N}} \qquad R > 3\Delta_{\mathcal{N}} + 7\Omega_{\mathcal{N}} + \frac{5}{2}\epsilon_{\mathcal{N}},$$

the split set $\mathcal{S}$ *can be reconstructed in polynomial time together with weight estimates* $\hat{w} : \mathcal{S} \to (0, +\infty)$ *satisfying* $|\hat{w}(S) - w(S)| < 2\tau$.

0	3	8	1	10	8
	0	9	2	11	9
		0	7	8	6
			0	9	7
				0	4
					0



Main results

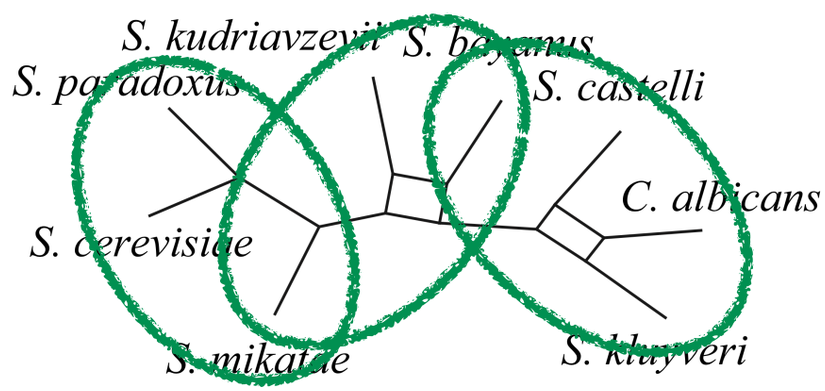
Theorem (R.-Wang'18)

Suppose $\mathcal{N} = (\mathcal{X}, \mathcal{S}, w)$ *is a circular network. Given a* (τ, R)*-distorted metric for* $\mathcal{N}$ *with*

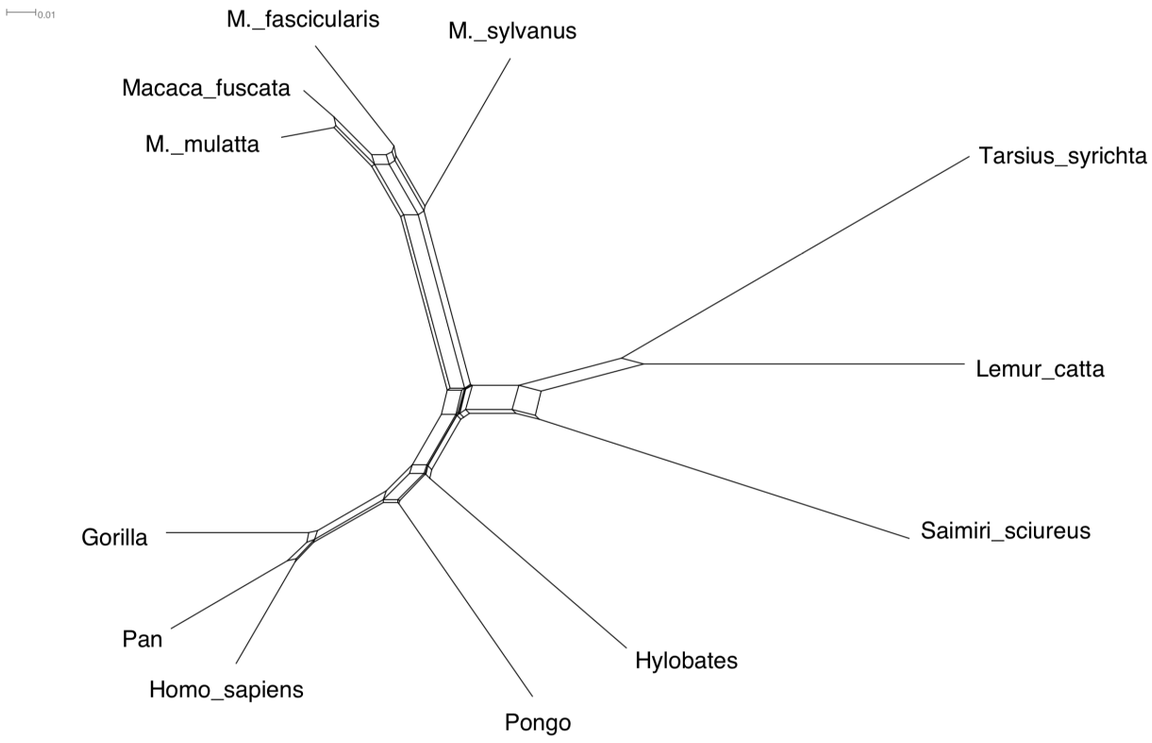
$$\tau < \frac{1}{4}\epsilon_{\mathcal{N}} \qquad R > 3\Delta_{\mathcal{N}} + 7\Omega_{\mathcal{N}} + \frac{5}{2}\epsilon_{\mathcal{N}},$$

the split set $\mathcal{S}$ *can be reconstructed in polynomial time together with weight estimates* $\hat{w} : \mathcal{S} \to (0, +\infty)$ *satisfying* $|\hat{w}(S) - w(S)| < 2\tau$.

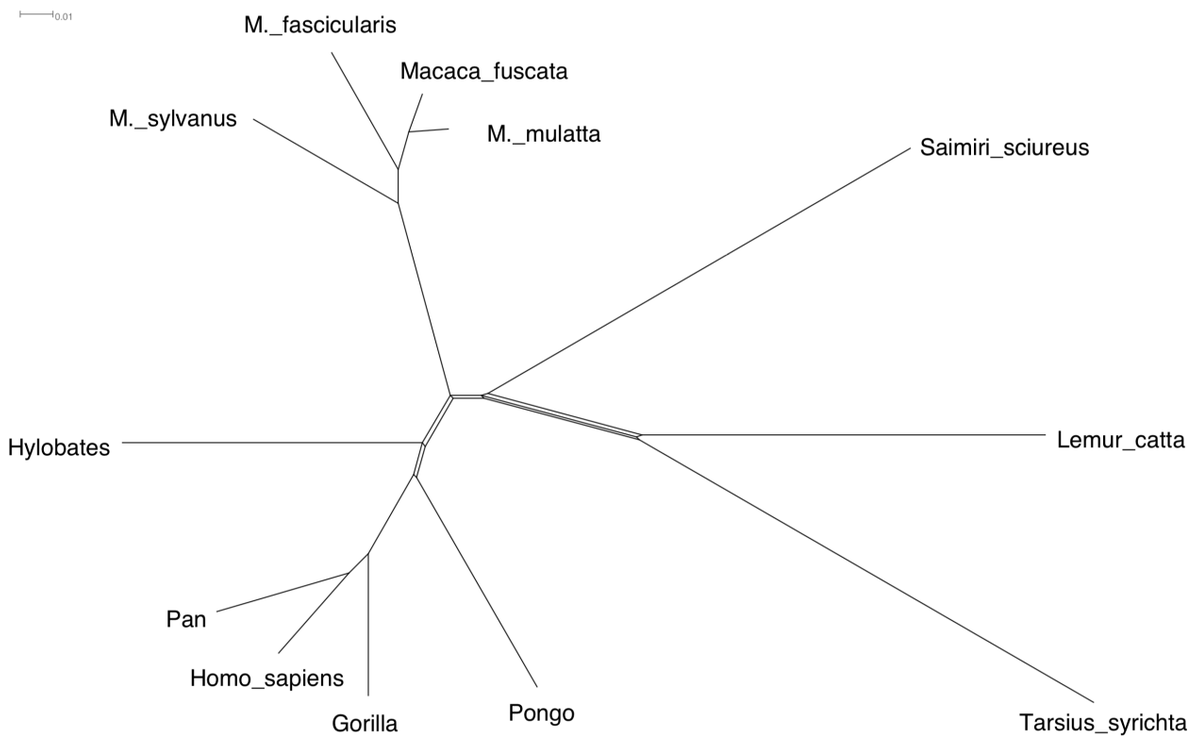
0	3	8	1	10	8
	0	9	2	11	9
		0	7	8	6
			0	9	7
				0	4
					0



Example



Neighbor-Net



Our method

Thanks

Work supported by:

SIMONS FOUNDATION

NSF

ALFRED P. SLOAN FOUNDATION 1934

For more details:
http://www.math.wisc.edu/~roch/