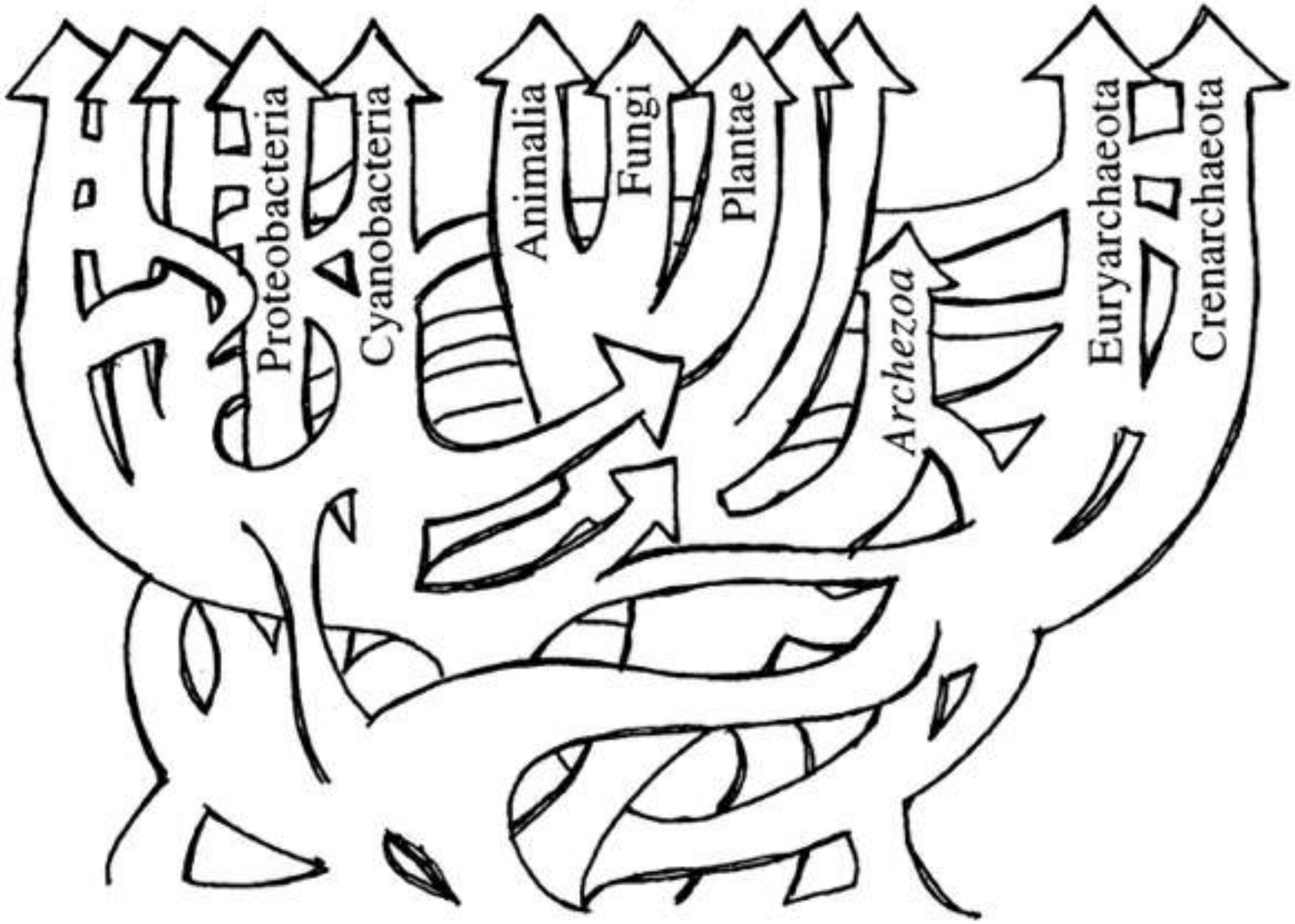
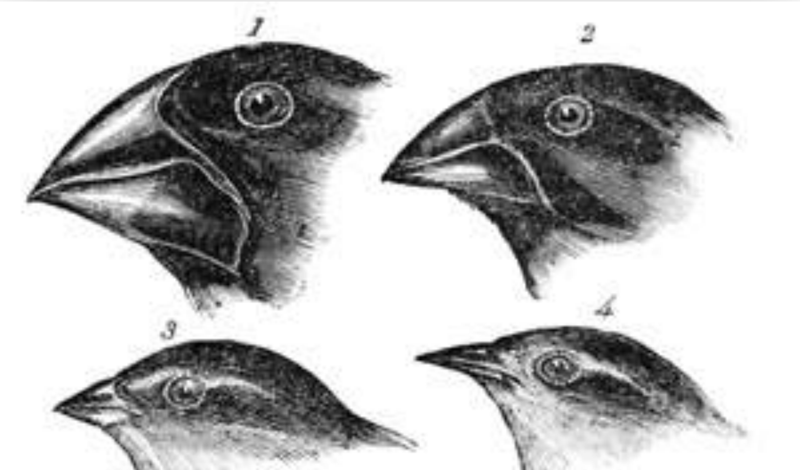


Mathematics of the Tree of Life

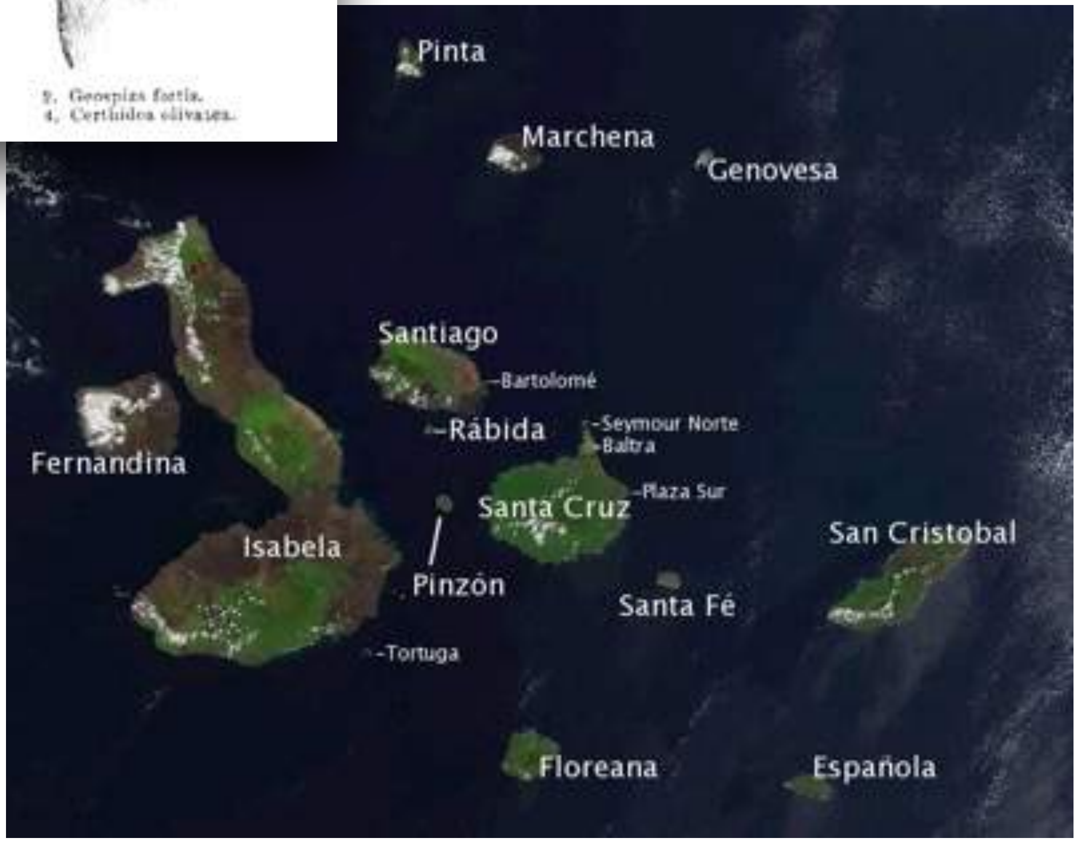
From Genomes to Phylogenetic Trees and Beyond

Sébastien Roch
Department of Mathematics
UW-Madison

Darwin’s finches

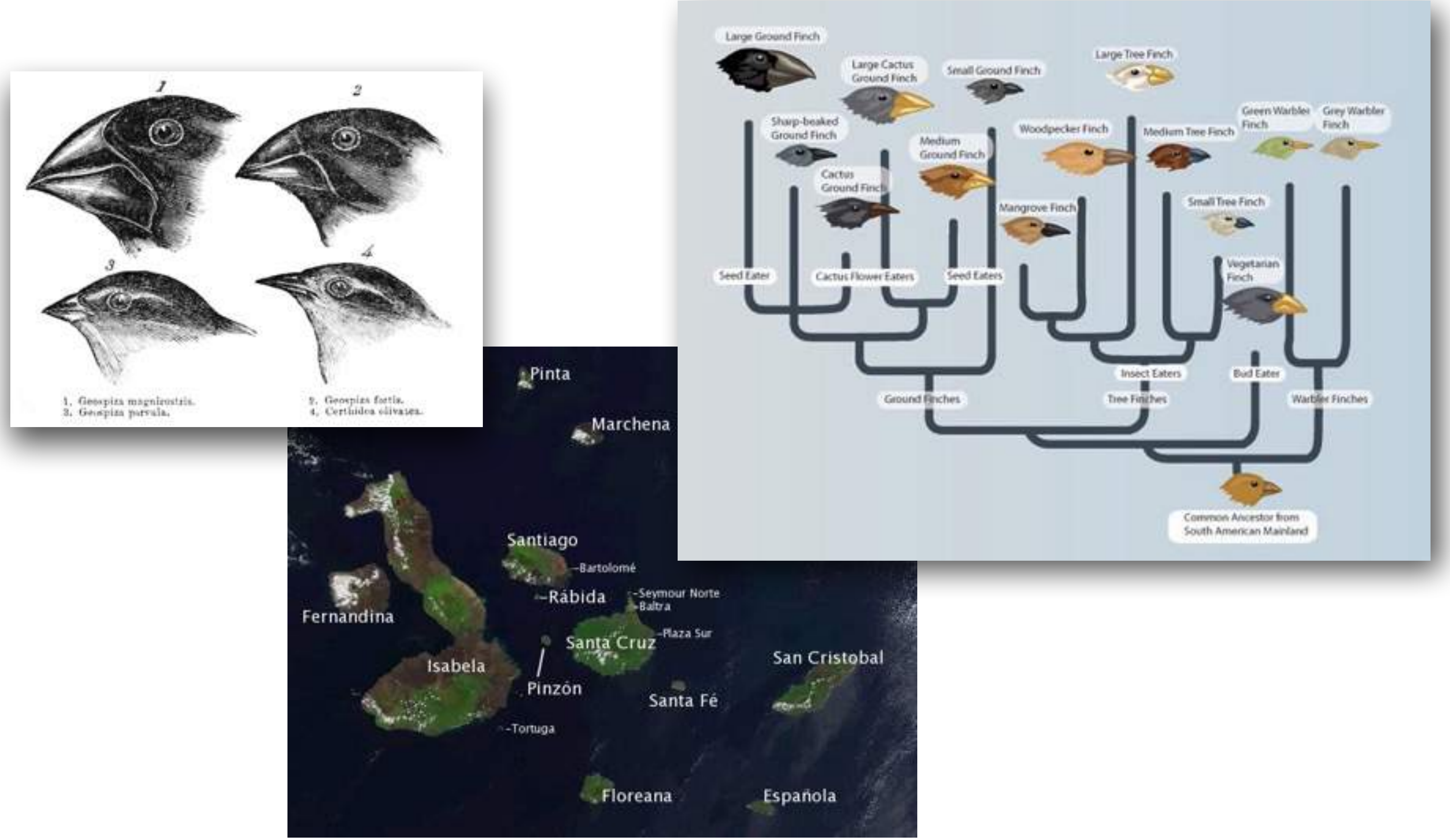


1. Geospiza magnirostris.
2. Geospiza fortis.
3. Geospiza parvula.
4. Certhidea olivacea.



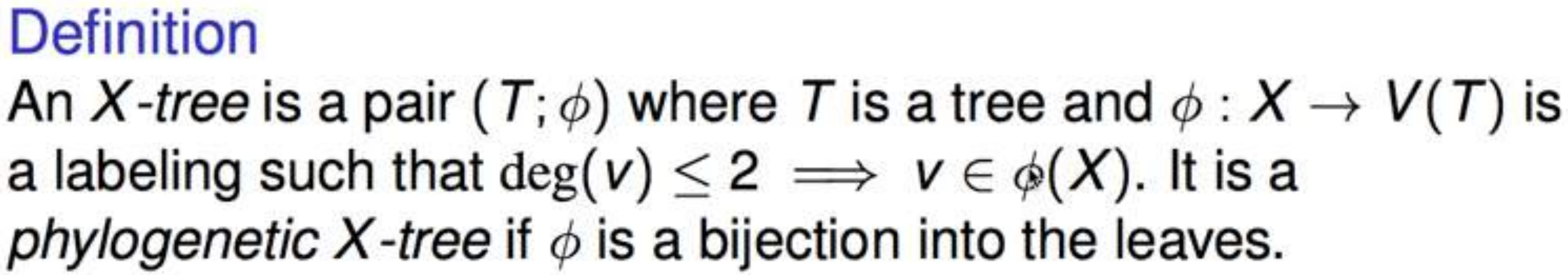
Right figure by Nature Education

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.

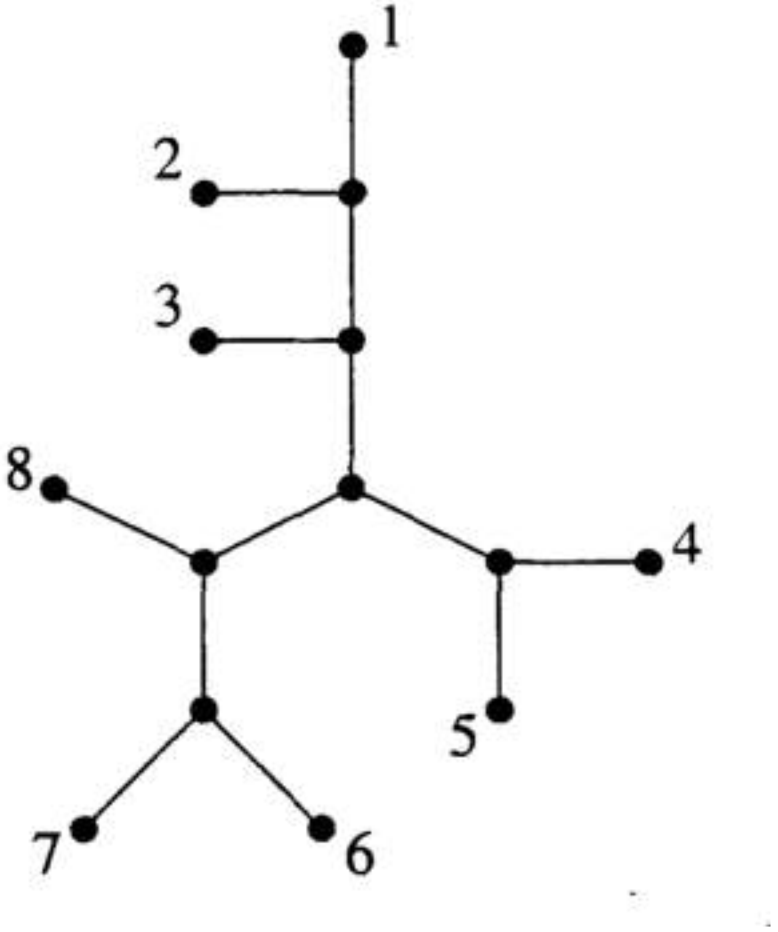


Figure from Semple & Steel, Phylogenetics (2003).

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 $\phi_2 = \psi \circ \phi_1$.

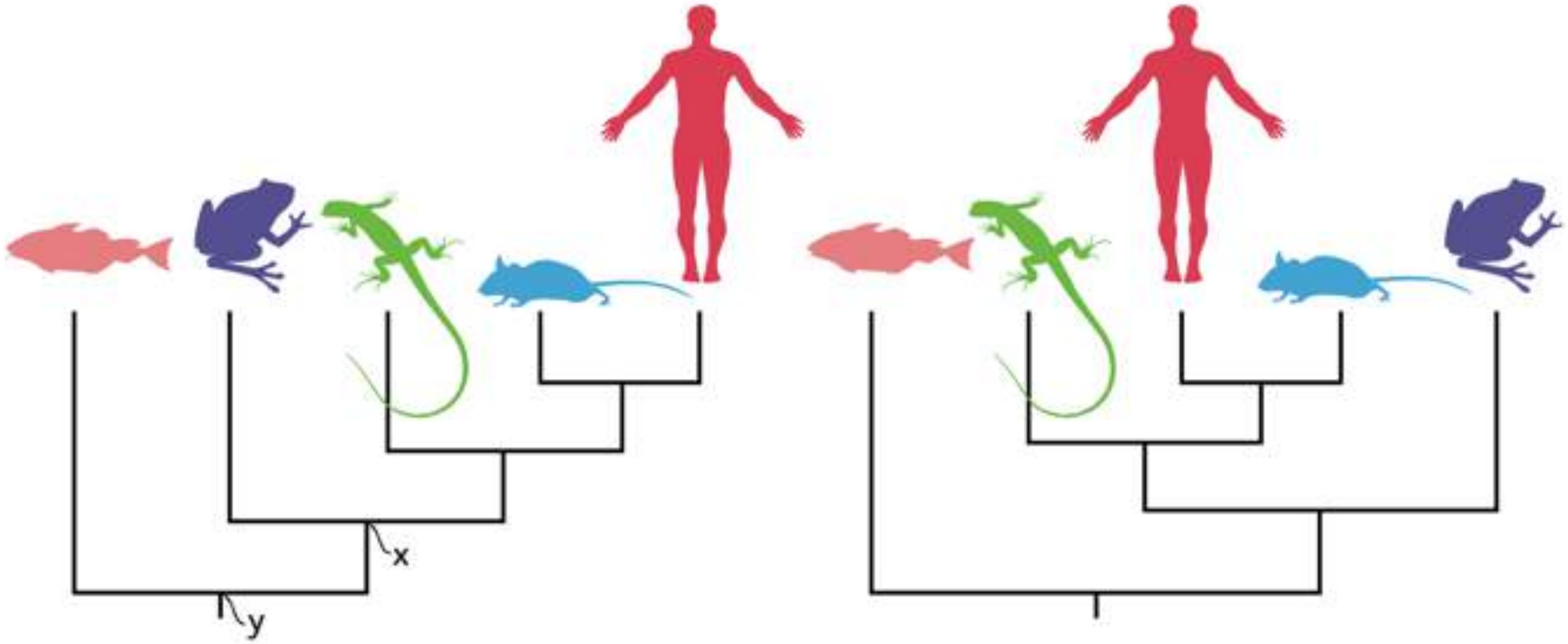


Figure from Baum et al. (2003). Science 310 (5750):979-980.

So how is the Tree of Life inferred?

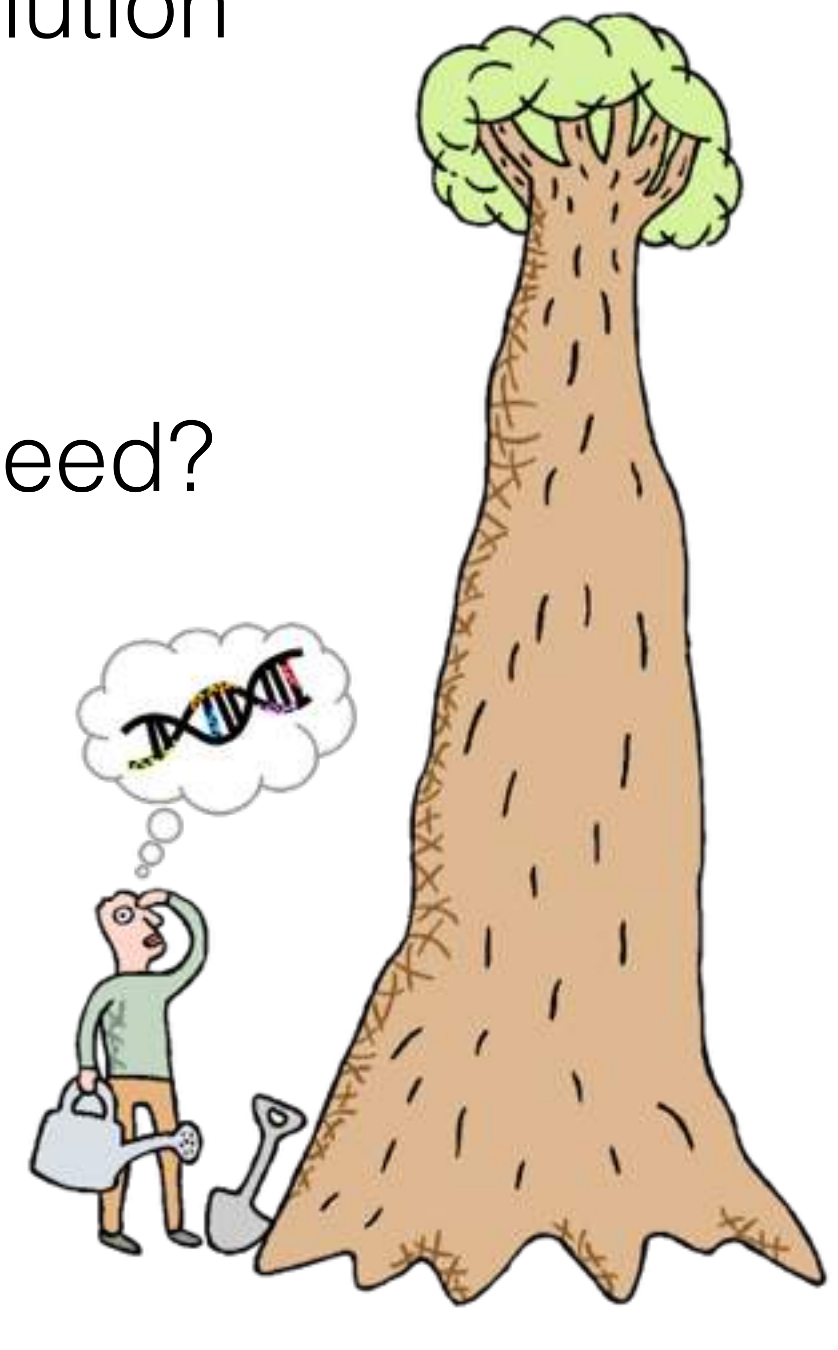
I. From Darwin’s finches to HIV evolution

II. Pre-genomics era

III. Transition: How much data do I need?

IV. More data, more problems

V. Is the Tree of Life even a tree?

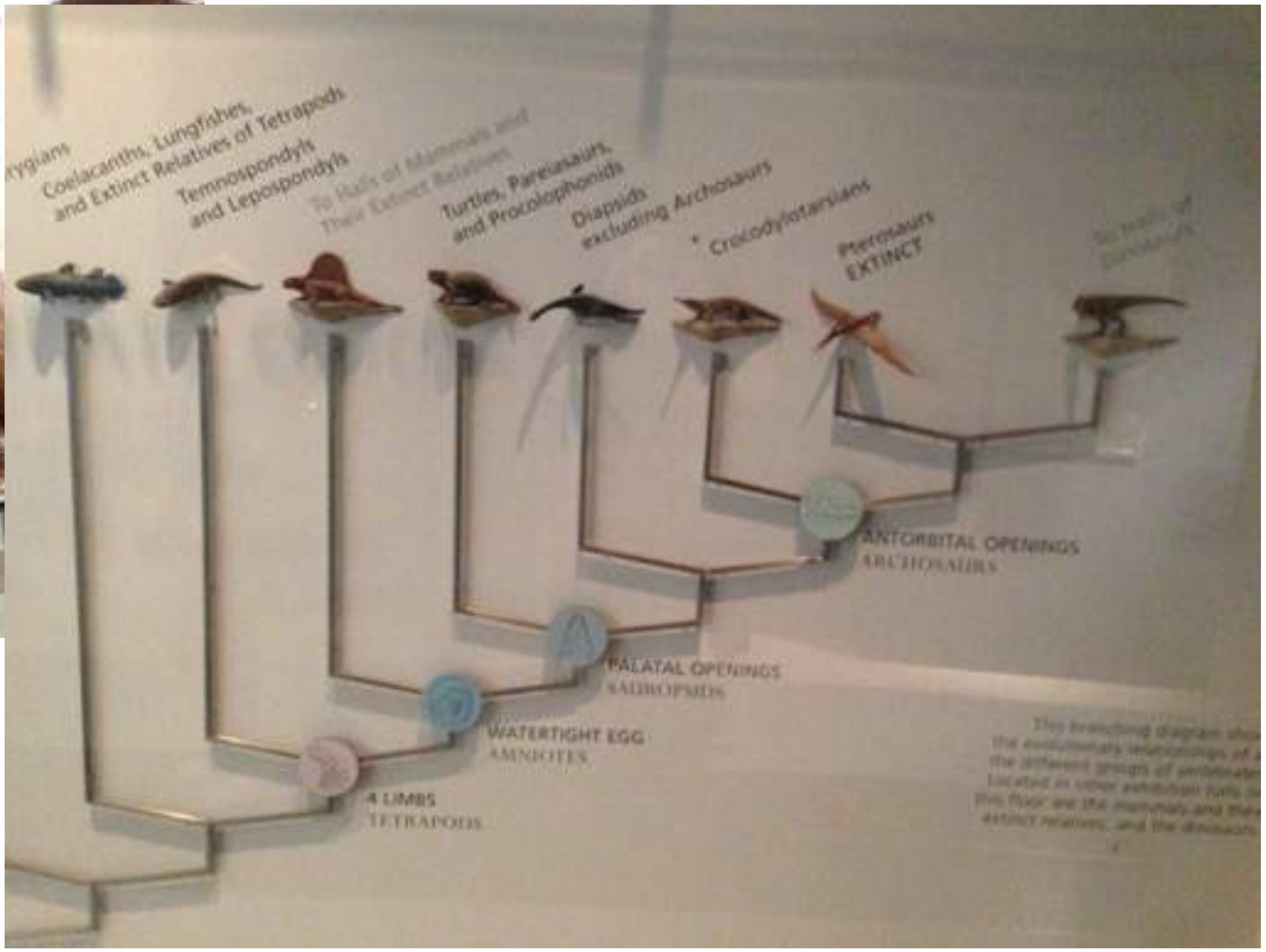


Pre-genomics era

Strolling down the Tree of Life

Strolling down the Tree of Life

Strolling down the Tree of Life


rygians
Coelacanths, Lungfishes, and Extinct Relatives of Tetrapods
Temnospondyls and Lepospondyls
To Halls of Mammals and Their Extinct Relatives
Turtles, Pareiasaurs, and Procolophonids
Diapsids excluding Archosaurs
* Crocodylotarsians
Pterosaurs EXTINCT
To Halls of Dinosaurs
ANTORBITAL OPENINGS
ARCHOSAURS
PALATAL OPENINGS
SAUROPSIDS
WATERTIGHT EGG
AMNIOTES
4 LIMBS
TETRAPODS
This branching diagram sho
the evolutionary relationships of a
the different groups of vertebrat
Located in other exhibition halls o
this floor are the mammals and the
extinct relatives, and the dinosaurs

Strolling down the Tree of Life

rygians
Coelacanths, Lungfishes, and Extinct Relatives of Tetrapods
Temnospondyls and Lepospondyls
To Halls of Mammals and Their Extinct Relatives
Turtles, Pareiasaurs, and Procolophonids
Diapsids excluding Archosaurs
Crocodylotarsians
Pterosaurs EXTINCT
ANTORBITAL
ARCHOSAU
PALATAL OPENINGS
SAUROPSIDS
WATERTIGHT EGG
AMNIOTES
4 LIMBS
TETRAPODS
JAWS

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.

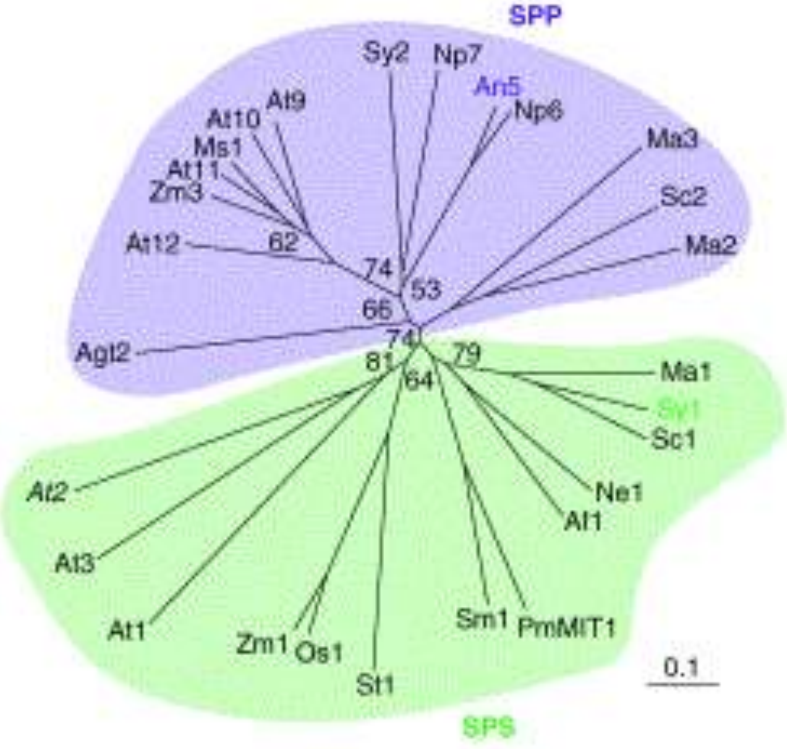


Figure from Salerno, Curatti, Trends Plant Sci. 2003 Feb;8(2):63-9.

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.

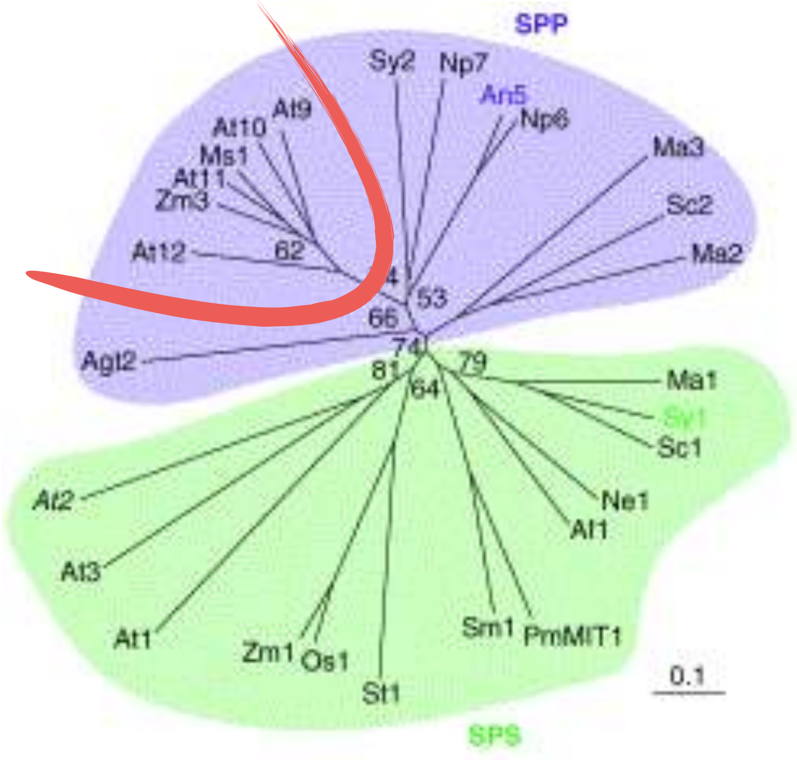
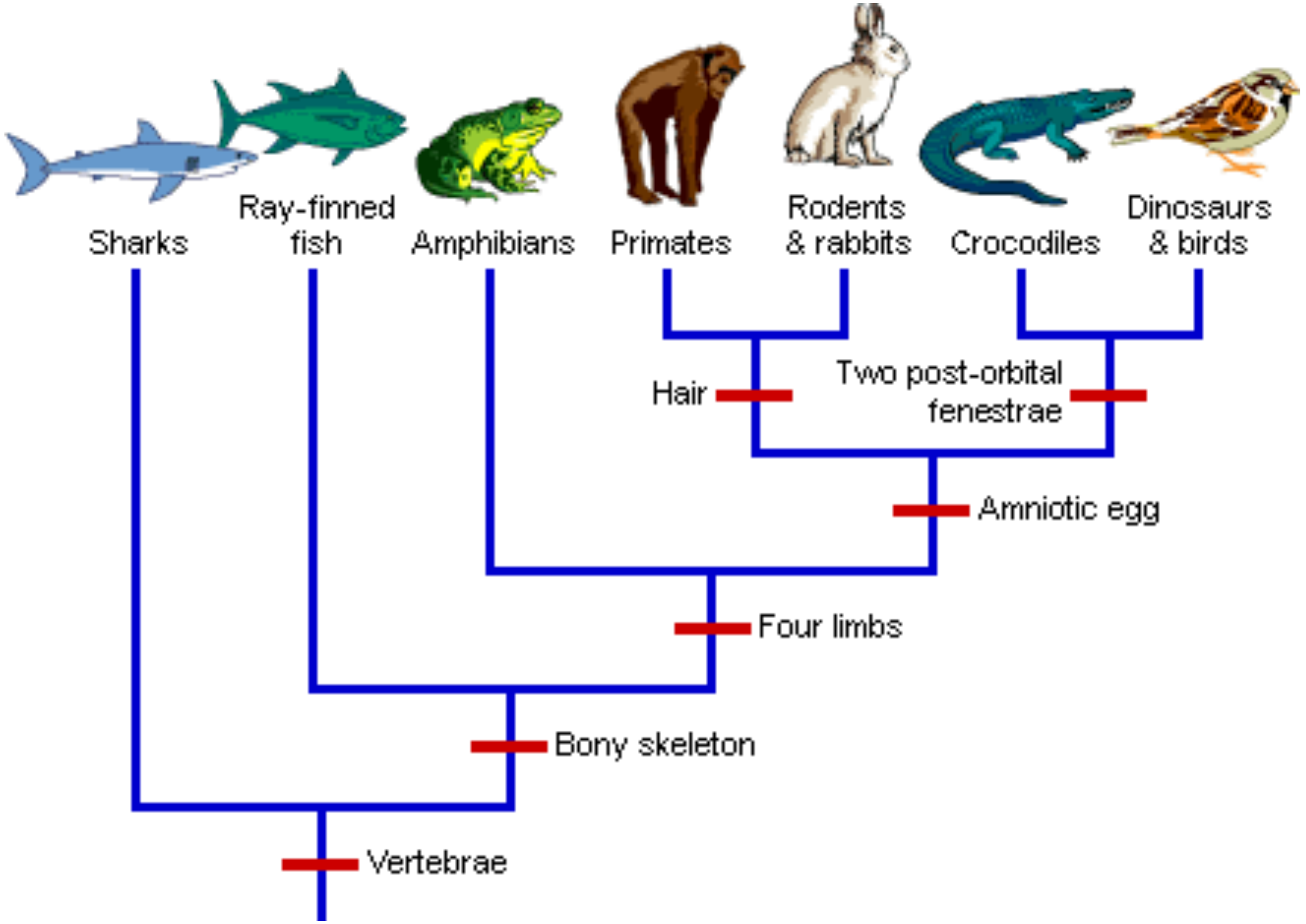


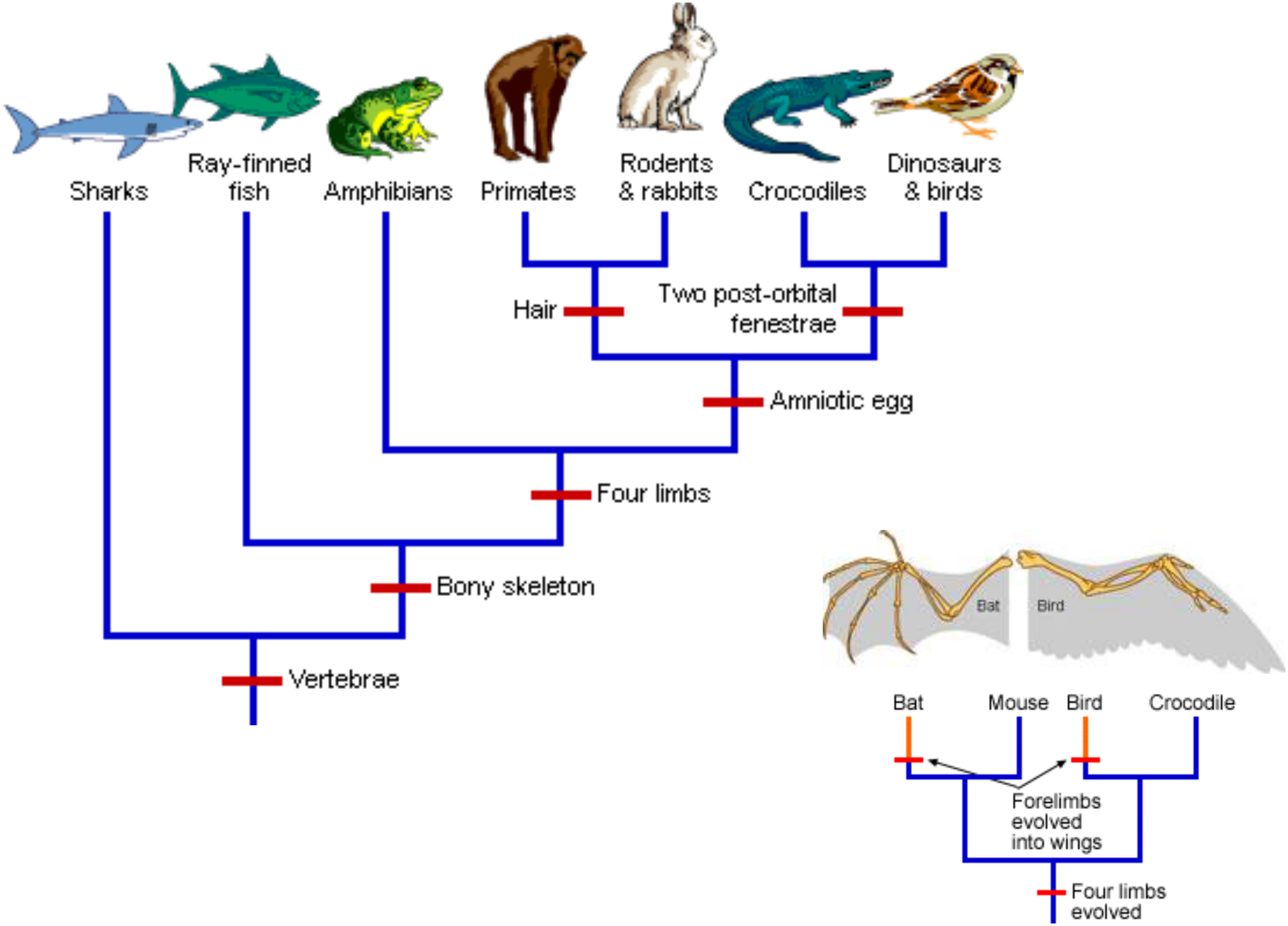
Figure from Salerno, Curatti, Trends Plant Sci. 2003 Feb;8(2):63-9.

Synapomorphies & homoplasies



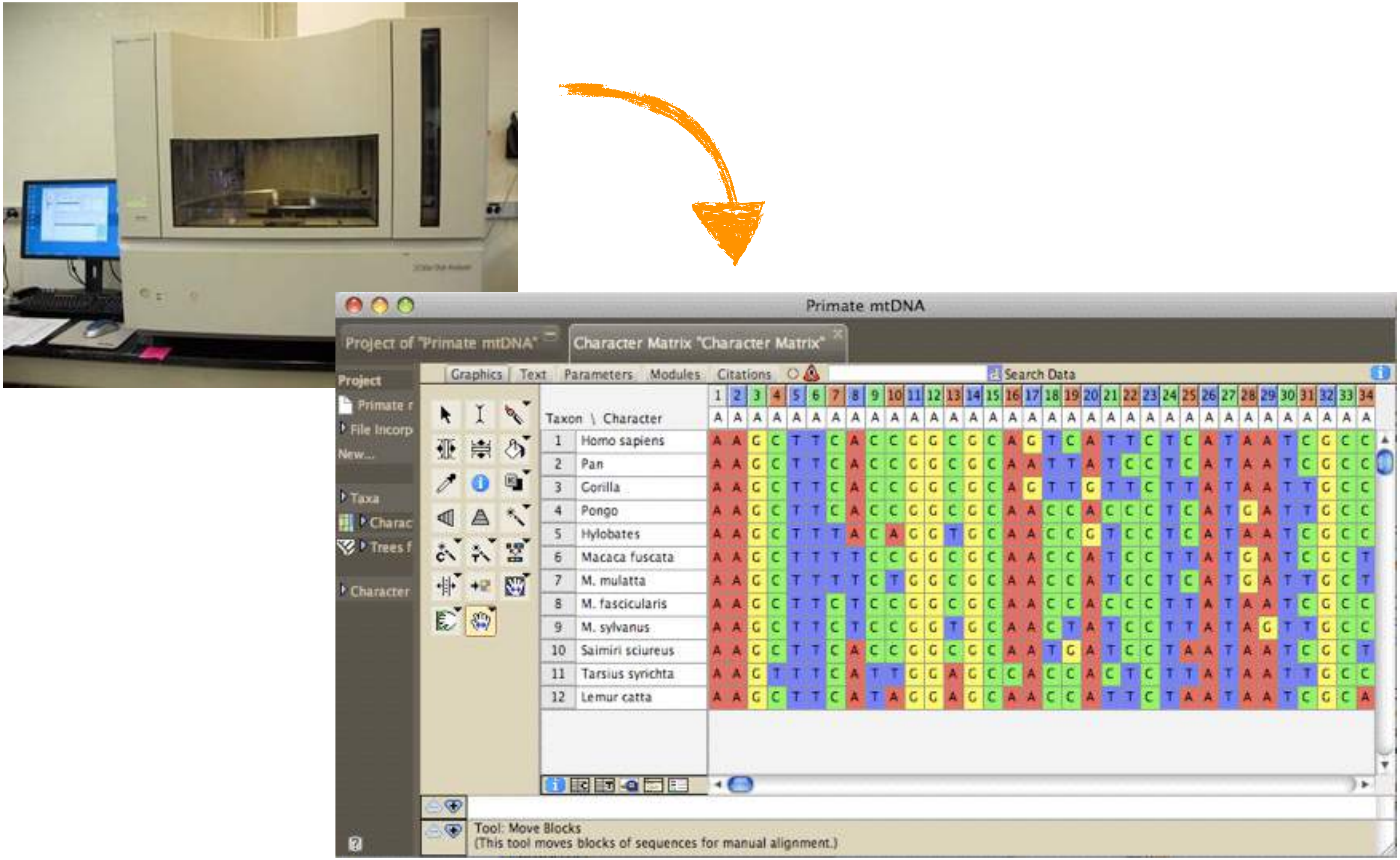
Figures by University of California Museum of Paleontology's Understanding Evolution

Synapomorphies & homoplasies



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

Molecular systematics

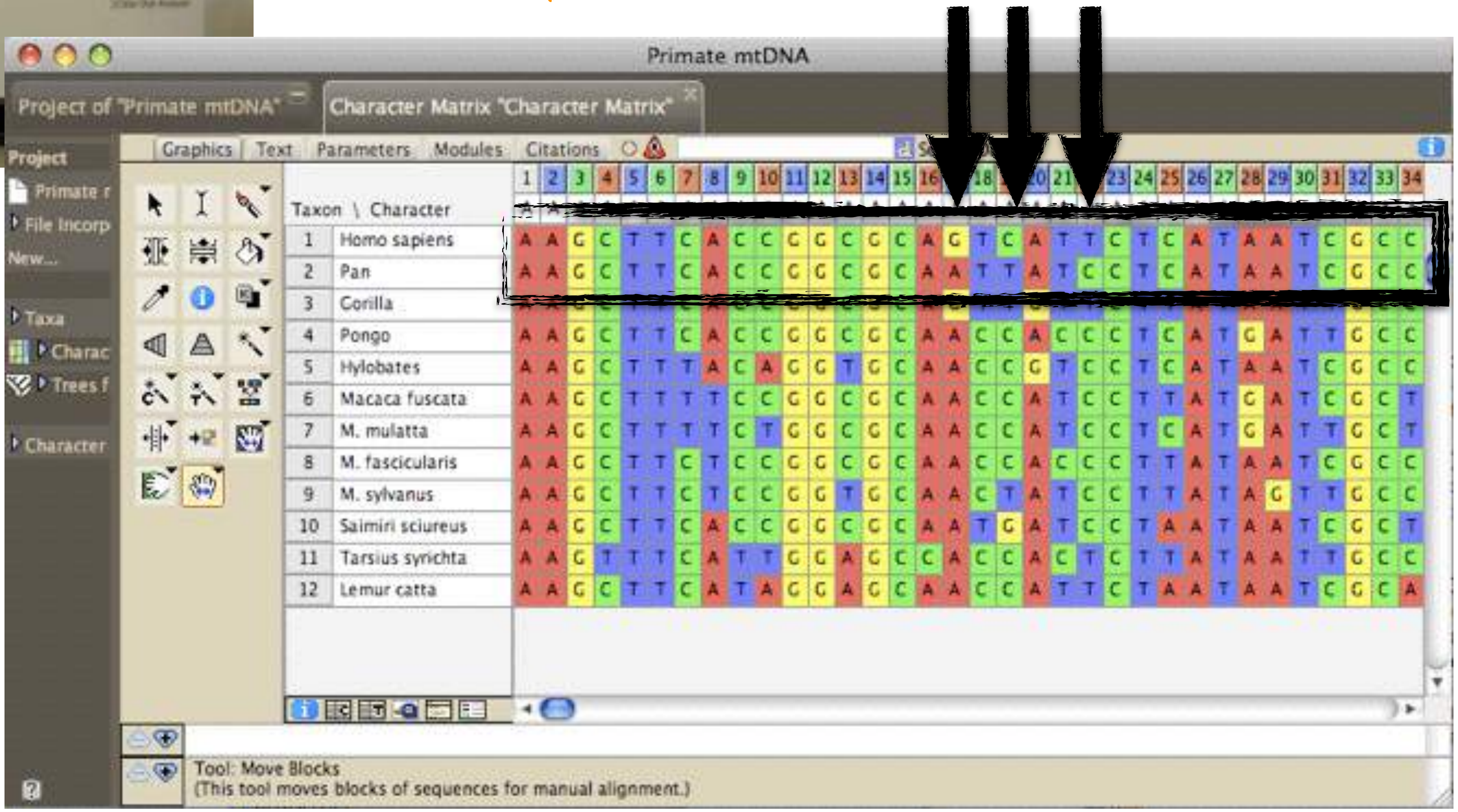


	Taxon \ Character	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
		A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
1	Homo sapiens	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	C	A	T	T	C	T	C	A	T	A	A	T	C	G	C	C
2	Pan	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	T	A	T	C	C	T	C	A	T	A	A	T	C	G	C	C
3	Gorilla	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	T	G	T	T	C	T	T	A	T	A	A	T	T	G	C	C
4	Pongo	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	C	A	T	G	A	T	T	G	C	C
5	Hylobates	A	A	G	C	T	T	T	A	C	A	G	G	T	G	C	A	A	C	C	G	T	C	C	T	C	A	T	A	A	T	C	G	C	C
6	Macaca fuscata	A	A	G	C	T	T	T	T	C	C	G	G	C	G	C	A	A	C	C	A	T	C	C	T	T	A	T	G	A	T	C	G	C	T
7	M. mulatta	A	A	G	C	T	T	T	T	C	T	G	G	C	G	C	A	A	C	C	A	T	C	C	T	C	A	T	G	A	T	T	G	C	T
8	M. fascicularis	A	A	G	C	T	T	C	T	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	T	A	T	A	A	T	C	G	C	C
9	M. sylvanus	A	A	G	C	T	T	C	T	C	C	G	G	T	G	C	A	A	C	T	A	T	C	C	T	T	A	T	A	G	T	T	G	C	C
10	Saimiri sciureus	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	G	A	T	C	C	T	A	A	T	A	A	T	C	G	C	T
11	Tarsius syrichta	A	A	G	T	T	T	C	A	T	T	G	G	A	G	C	C	A	C	C	A	C	T	C	T	T	A	T	A	A	T	T	G	C	C
12	Lemur catta	A	A	G	C	T	T	C	A	T	A	G	G	A	G	C	A	A	C	C	A	T	T	C	T	A	A	T	A	A	T	C	G	C	A

Tool: Move Blocks
(This tool moves blocks of sequences for manual alignment.)

Snapshot from Mesquite

Molecular systematics

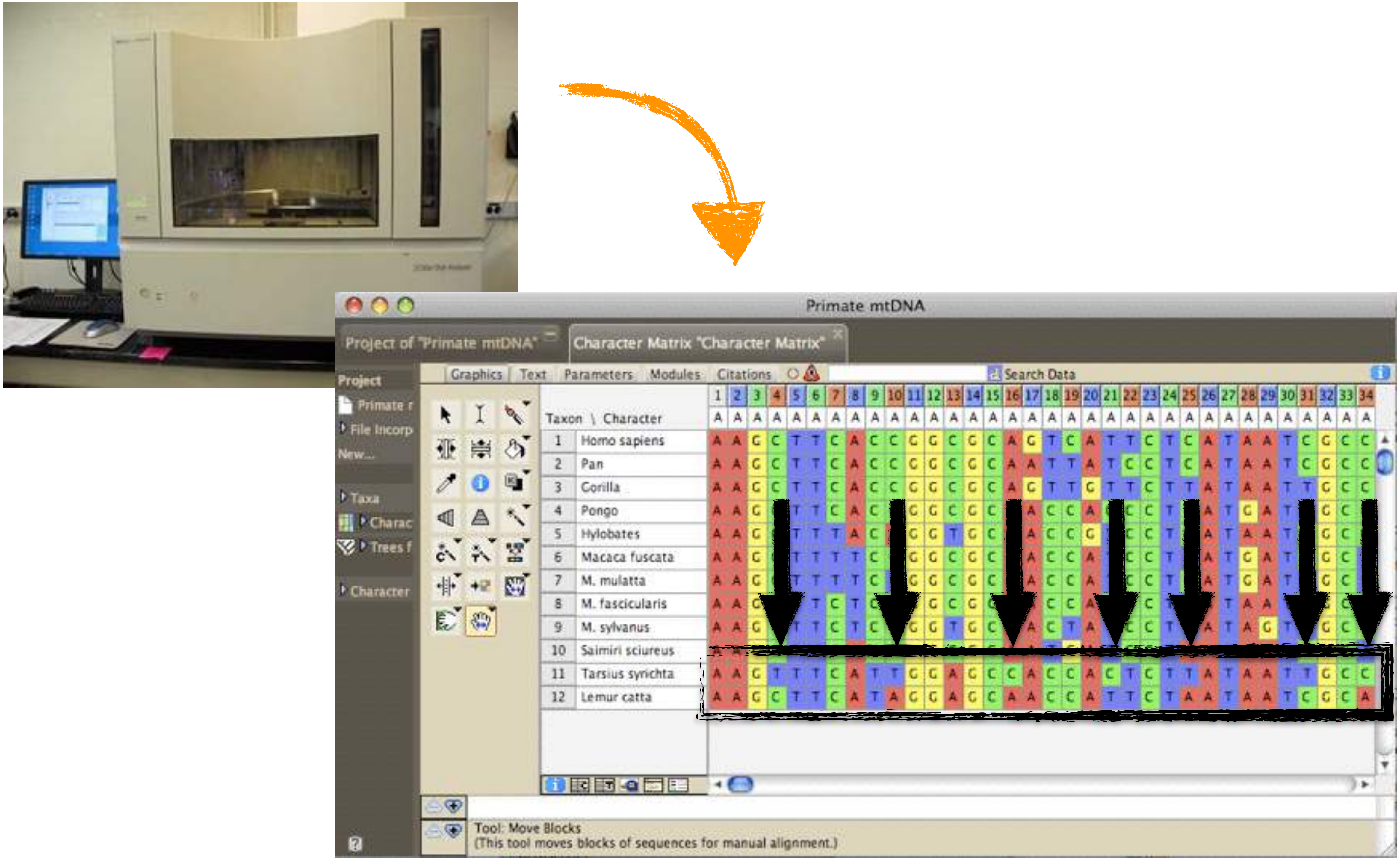


	Taxon \ Character	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	Homo sapiens	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	C	A	T	T	C	T	C	A	T	A	A	T	C	G	C	C
2	Pan	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	T	A	T	C	C	T	C	A	T	A	A	T	C	G	C	C
3	Gorilla	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	T	G	T	T	C	T	T	A	T	A	A	T	T	G	C	C
4	Pongo	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	C	A	T	G	A	T	T	G	C	C
5	Hylobates	A	A	G	C	T	T	T	A	C	A	G	G	T	G	C	A	A	C	C	G	T	C	C	T	C	A	T	A	A	T	C	G	C	C
6	Macaca fuscata	A	A	G	C	T	T	T	T	C	C	G	G	C	G	C	A	A	C	C	A	T	C	C	T	T	A	T	G	A	T	C	G	C	T
7	M. mulatta	A	A	G	C	T	T	T	T	C	T	G	G	C	G	C	A	A	C	C	A	T	C	C	T	C	A	T	G	A	T	T	G	C	T
8	M. fascicularis	A	A	G	C	T	T	C	T	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	T	A	T	A	A	T	C	G	C	C
9	M. sylvanus	A	A	G	C	T	T	C	T	C	C	G	G	T	G	C	A	A	C	T	A	T	C	C	T	T	A	T	A	G	T	T	G	C	C
10	Saimiri sciureus	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	G	A	T	C	C	T	A	A	T	A	A	T	C	G	C	T
11	Tarsius syrichta	A	A	G	T	T	T	C	A	T	T	G	G	A	G	C	C	A	C	C	A	C	T	C	T	T	A	T	A	A	T	T	G	C	C
12	Lemur catta	A	A	G	C	T	T	C	A	T	A	G	G	A	G	C	A	A	C	C	A	T	T	C	T	A	A	T	A	A	T	C	G	C	A

Tool: Move Blocks
(This tool moves blocks of sequences for manual alignment.)

Snapshot from Mesquite

Molecular systematics



	Taxon \ Character	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
		A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
1	Homo sapiens	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	C	A	T	T	C	T	C	A	T	A	A	T	C	G	C	C
2	Pan	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	T	A	T	C	C	T	C	A	T	A	A	T	C	G	C	C
3	Gorilla	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	T	G	T	T	C	T	T	A	T	A	A	T	T	G	C	C
4	Pongo	A	A	G	[illegible]	T	T	C	A	C	[illegible]	G	G	C	G	C	[illegible]	A	C	C	A	[illegible]	C	C	T	[illegible]	A	T	G	A	T	[illegible]	G	C	[illegible]
5	Hylobates	A	A	G	[illegible]	T	T	T	A	C	[illegible]	G	G	T	G	C	[illegible]	A	C	C	G	[illegible]	C	C	T	[illegible]	A	T	A	A	T	[illegible]	G	C	[illegible]
6	Macaca fuscata	A	A	G	[illegible]	T	T	T	T	C	[illegible]	G	G	C	G	C	[illegible]	A	C	C	A	[illegible]	C	C	T	[illegible]	A	T	G	A	T	[illegible]	G	C	[illegible]
7	M. mulatta	A	A	G	[illegible]	T	T	T	T	C	[illegible]	G	G	C	G	C	[illegible]	A	C	C	A	[illegible]	C	C	T	[illegible]	A	T	G	A	T	[illegible]	G	C	[illegible]
8	M. fascicularis	A	A	G	[illegible]	[illegible]	T	C	T	C	[illegible]	[illegible]	G	C	G	[illegible]	[illegible]	[illegible]	C	C	A	[illegible]	[illegible]	C	T	[illegible]	[illegible]	T	A	A	[illegible]	[illegible]	G	C	[illegible]
9	M. sylvanus	A	A	G	[illegible]	T	T	C	T	C	[illegible]	G	G	T	G	C	[illegible]	A	C	T	A	[illegible]	C	C	T	[illegible]	A	T	A	G	T	[illegible]	G	C	[illegible]
10	Saimiri sciureus	A	A	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11	Tarsius syrichta	A	A	G	T	T	T	C	A	T	T	G	G	A	G	C	C	A	C	C	A	C	T	C	T	T	A	T	A	A	T	T	G	C	C
12	Lemur catta	A	A	G	C	T	T	C	A	T	A	G	G	A	G	C	A	A	C	C	A	T	T	C	T	A	A	T	A	A	T	C	G	C	A

Tool: Move Blocks
(This tool moves blocks of sequences for manual alignment.)

Snapshot from Mesquite

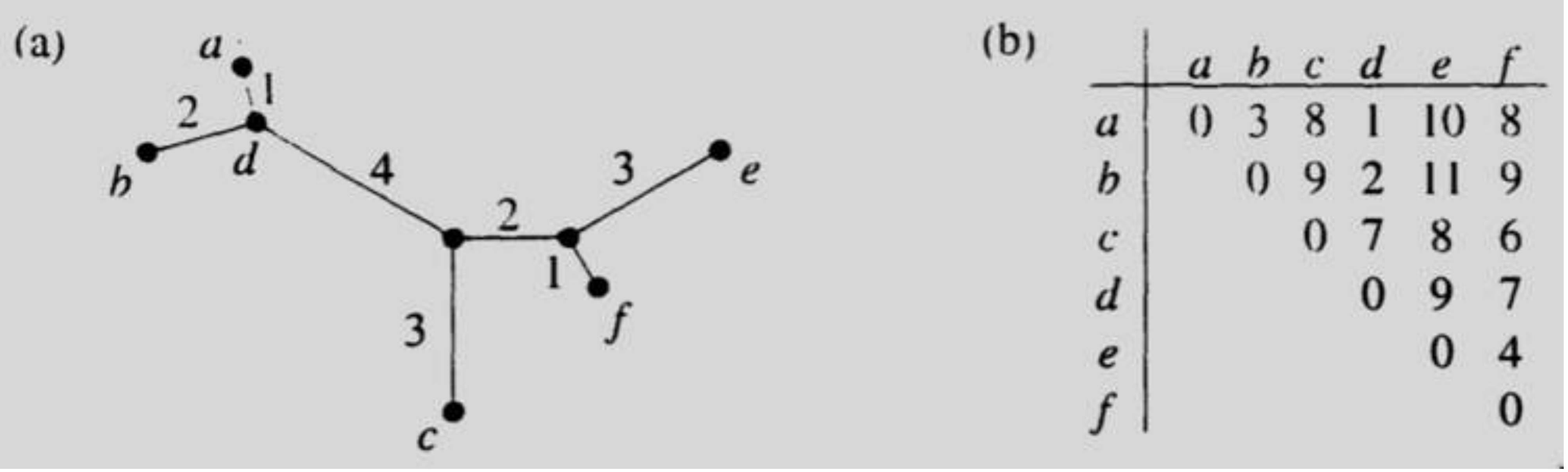
Tree metrics

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 (and efficiently computable).



	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).

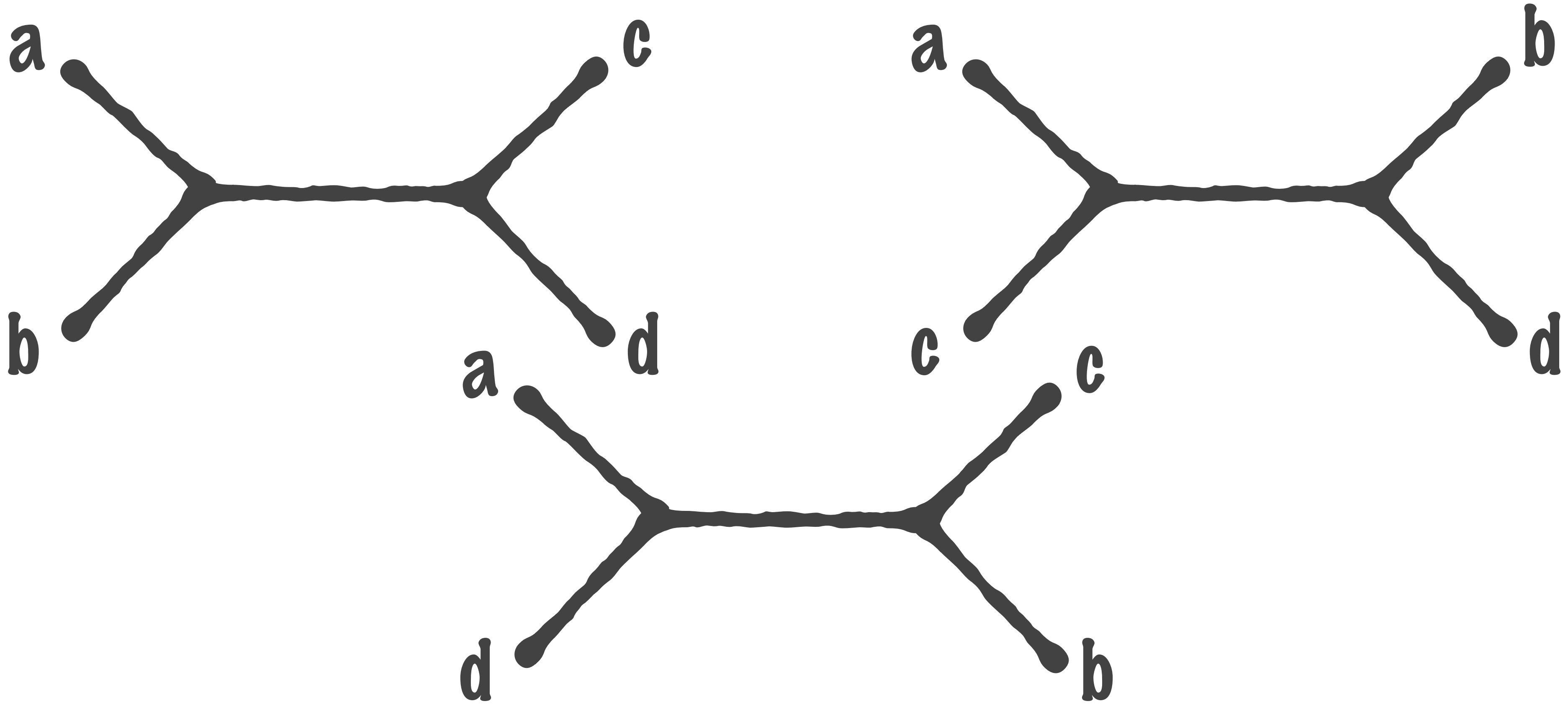
Tree metrics

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 (and efficiently computable).



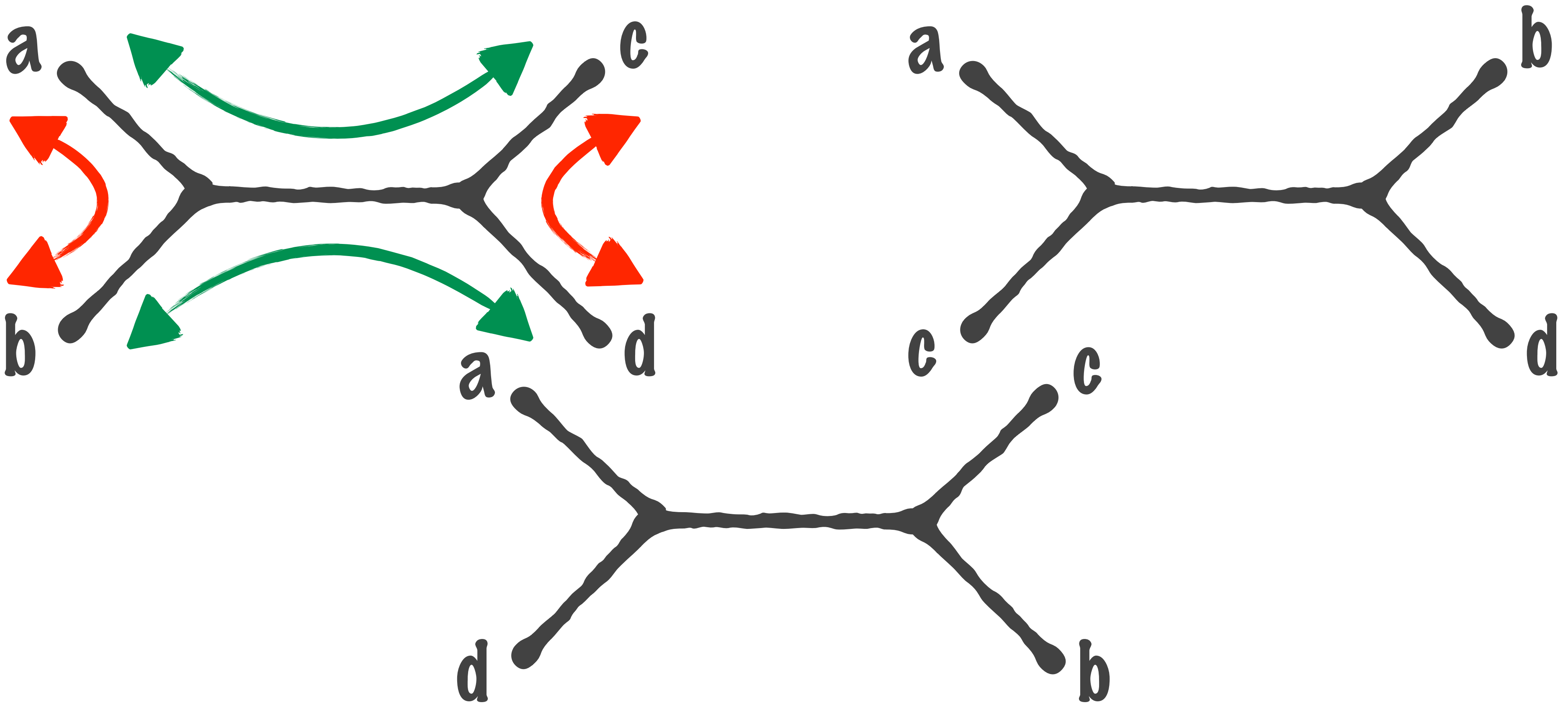
Tree metrics

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 (and efficiently computable).



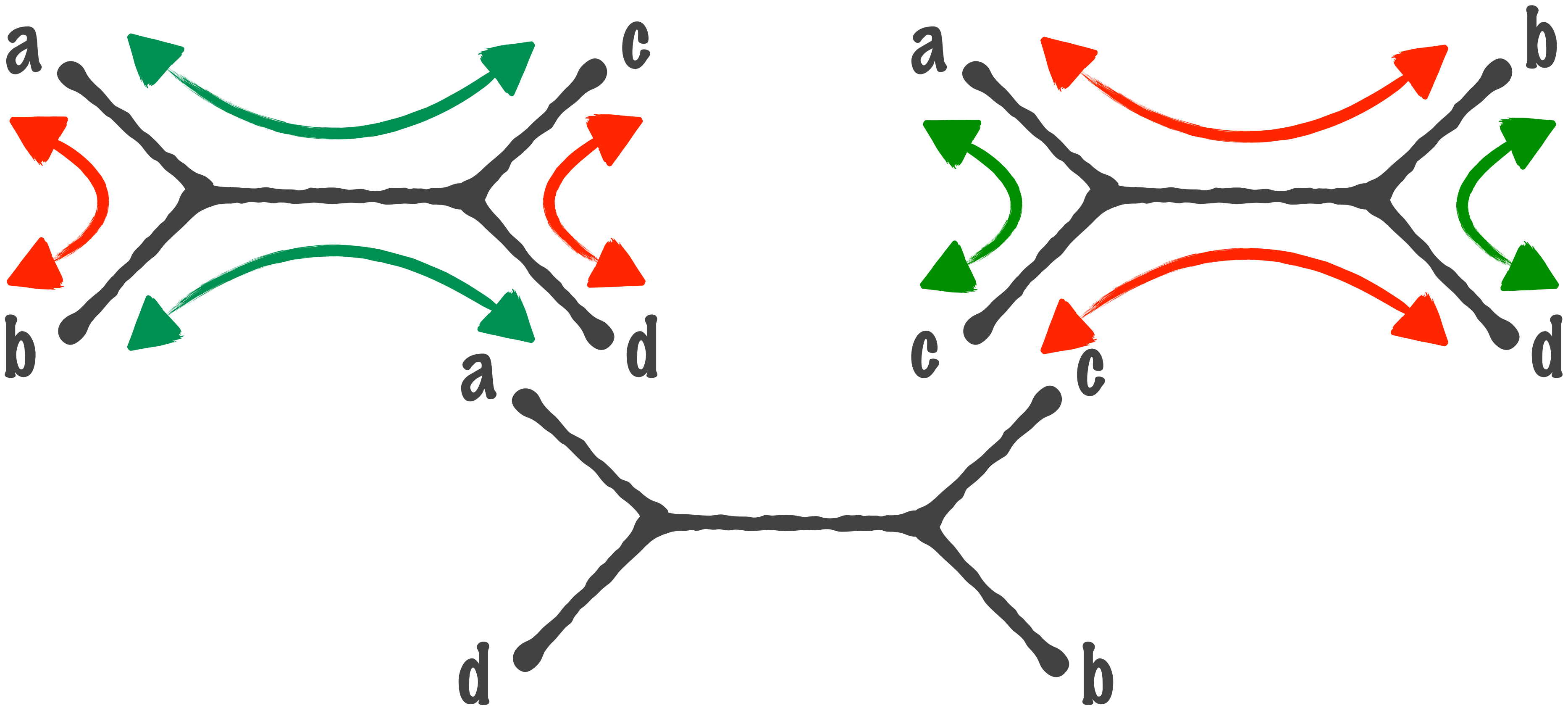
Tree metrics

Definition

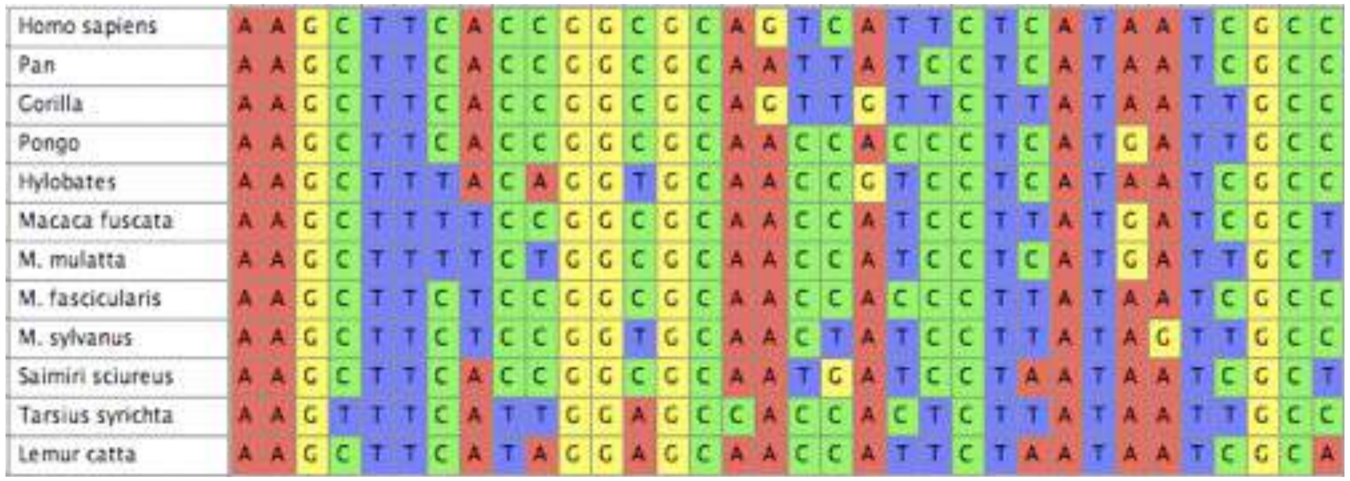
A function $\delta : X \times X \to \mathbb{R}$ is a *tree metric* if there is an X-tree $\mathcal{T} = (T; \phi)$ and a weighting $w : E(T) \to \mathbb{R}_+$ such that for all x, y

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

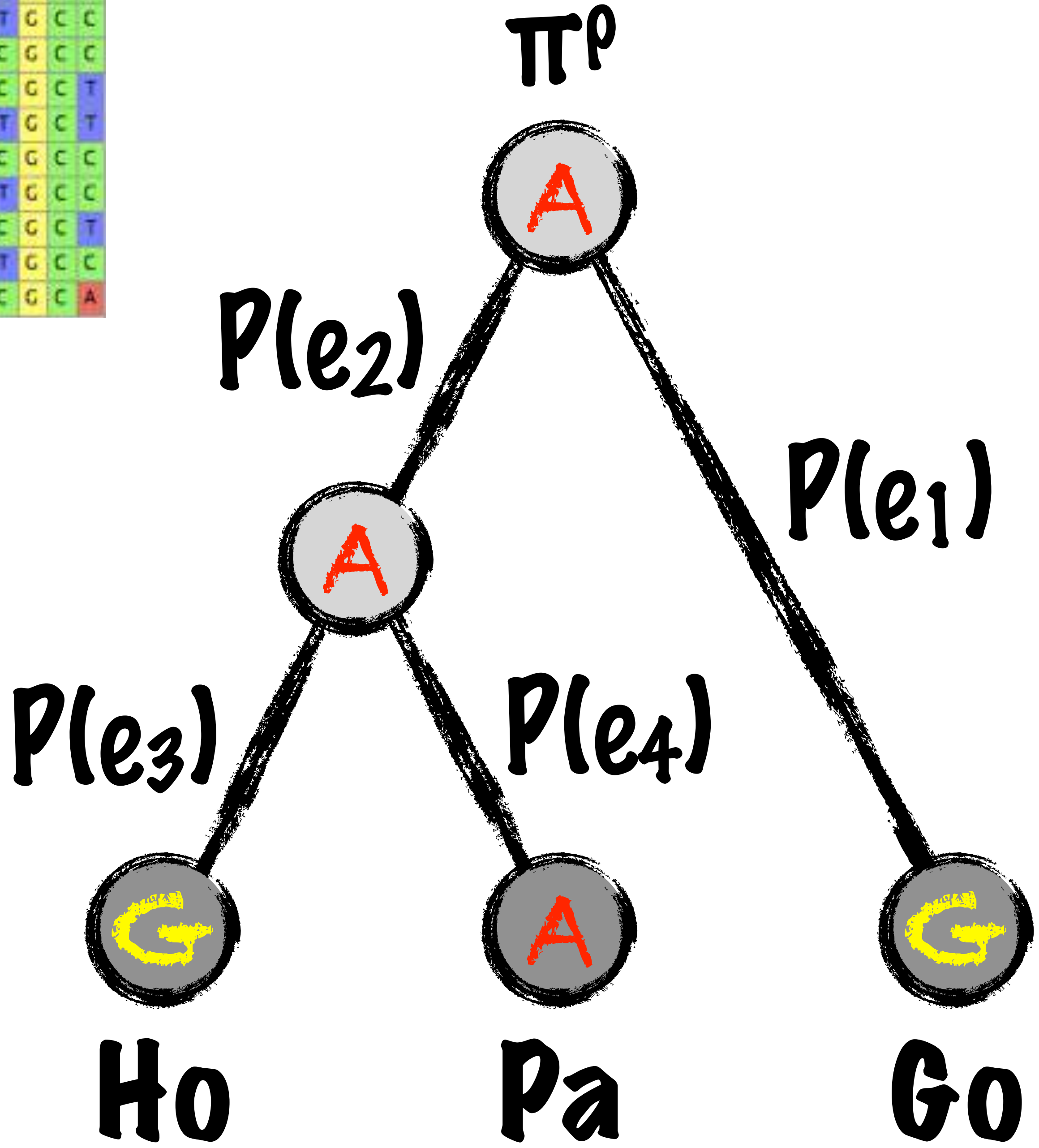
where $P(\mathcal{T}; x, y)$ is the unique path between $\phi(x)$ and $\phi(y)$. The *tree metric representation* $(\mathcal{T}; w)$ of δ is unique (and efficiently computable).



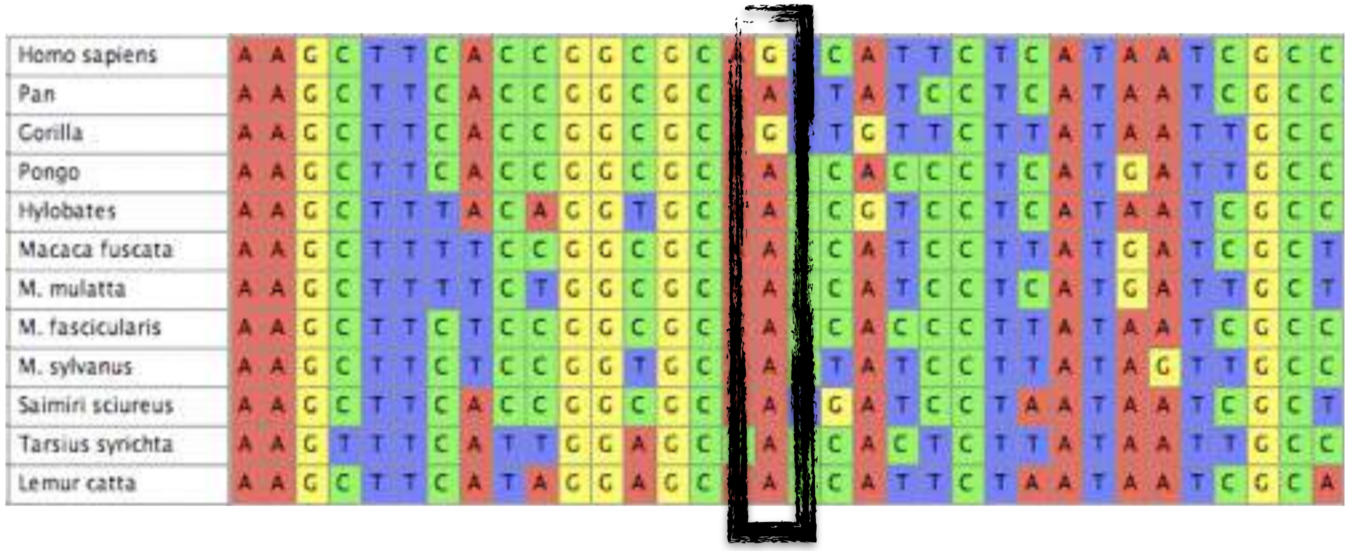
Markov process on a tree



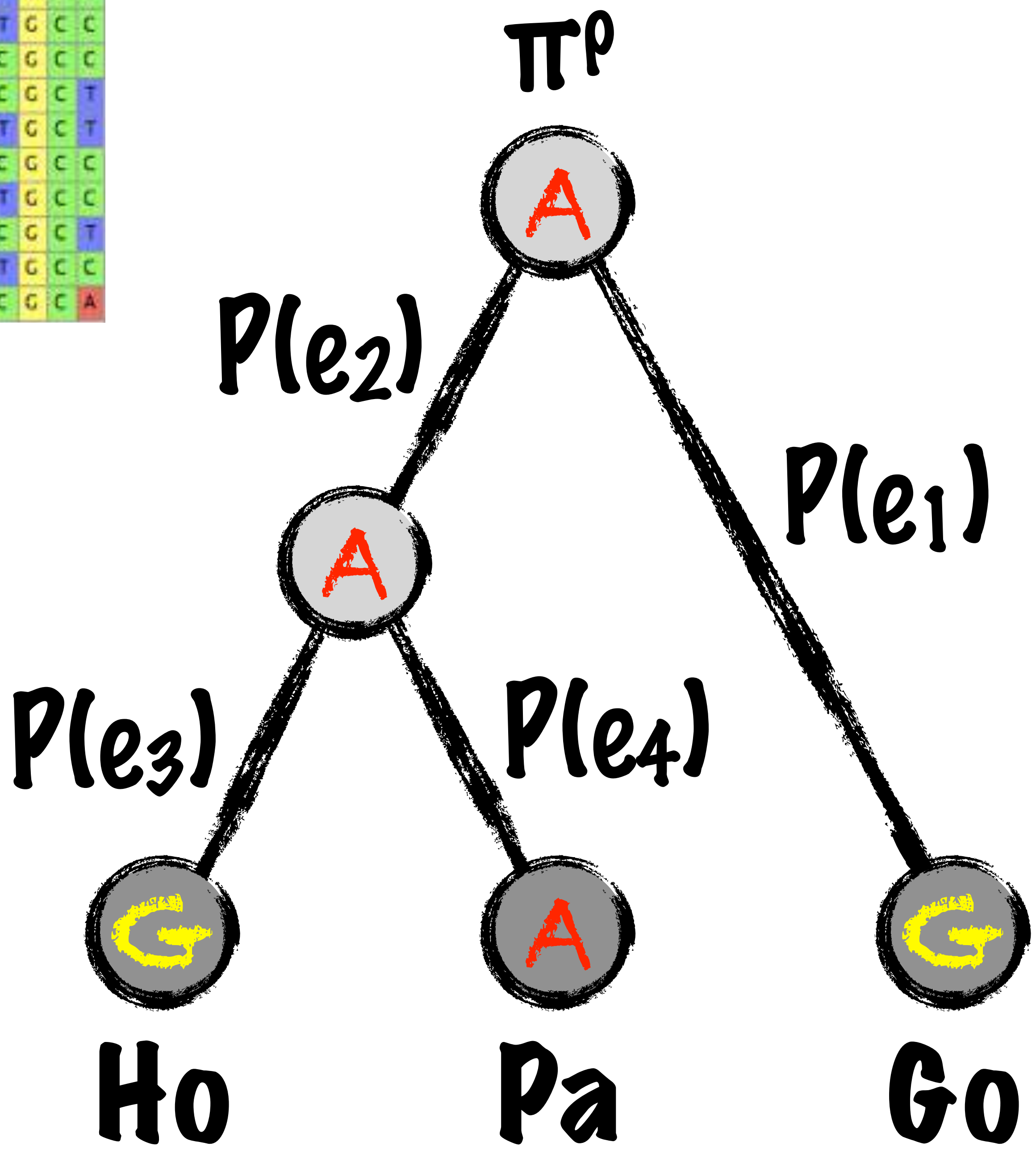
Homo sapiens	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	C	A	T	T	C	T	C	A	T	A	A	T	C	G	C	C
Pan	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	T	A	T	C	C	T	C	A	T	A	A	T	C	G	C	C
Gorilla	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	T	G	T	T	C	T	T	A	T	A	A	T	T	G	C	C
Pongo	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	C	A	T	G	A	T	T	G	C	C
Hylobates	A	A	G	C	T	T	T	A	C	A	G	G	T	G	C	A	A	C	C	G	T	C	C	T	C	A	T	A	A	T	C	G	C	C
Macaca fuscata	A	A	G	C	T	T	T	T	C	C	G	G	C	G	C	A	A	C	C	A	T	C	C	T	T	A	T	G	A	T	C	G	C	T
M. mulatta	A	A	G	C	T	T	T	T	C	T	G	G	C	G	C	A	A	C	C	A	T	C	C	T	C	A	T	G	A	T	T	G	C	T
M. fascicularis	A	A	G	C	T	T	C	T	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	T	A	T	A	A	T	C	G	C	C
M. sylvanus	A	A	G	C	T	T	C	T	C	C	G	G	T	G	C	A	A	C	T	A	T	C	C	T	T	A	T	A	G	T	T	G	C	C
Saimiri sciureus	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	G	A	T	C	C	T	A	A	T	A	A	T	C	G	C	T
Tarsius syrichta	A	A	G	T	T	T	C	A	T	T	G	G	A	G	C	C	A	C	C	A	C	T	C	T	T	A	T	A	A	T	T	G	C	C
Lemur catta	A	A	G	C	T	T	C	A	T	A	G	G	A	G	C	A	A	C	C	A	T	T	C	T	A	A	T	A	A	T	C	G	C	A



Markov process on a tree



Homo sapiens	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	[illegible]	C	A	T	T	C	T	C	A	T	A	A	T	C	G	C	C
Pan	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	[illegible]	A	[illegible]	T	A	T	C	C	T	C	A	T	A	A	T	C	G	C	C
Gorilla	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	[illegible]	G	[illegible]	T	G	T	T	C	T	T	A	T	A	A	T	T	G	C	C
Pongo	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	[illegible]	A	[illegible]	C	A	C	C	C	T	C	A	T	G	A	T	T	G	C	C
Hylobates	A	A	G	C	T	T	T	A	C	A	G	G	T	G	C	[illegible]	A	[illegible]	C	G	T	C	C	T	C	A	T	A	A	T	C	G	C	C
Macaca fuscata	A	A	G	C	T	T	T	T	C	C	G	G	C	G	C	[illegible]	A	[illegible]	C	A	T	C	C	T	T	A	T	G	A	T	C	G	C	T
M. mulatta	A	A	G	C	T	T	T	T	C	T	G	G	C	G	C	[illegible]	A	[illegible]	C	A	T	C	C	T	C	A	T	G	A	T	T	G	C	T
M. fascicularis	A	A	G	C	T	T	C	T	C	C	G	G	C	G	C	[illegible]	A	[illegible]	C	A	C	C	C	T	T	A	T	A	A	T	C	G	C	C
M. sylvanus	A	A	G	C	T	T	C	T	C	C	G	G	T	G	C	[illegible]	A	[illegible]	T	A	T	C	C	T	T	A	T	A	G	T	T	G	C	C
Saimiri sciureus	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	[illegible]	A	[illegible]	G	A	T	C	C	T	A	A	T	A	A	T	C	G	C	T
Tarsius syrichta	A	A	G	T	T	T	C	A	T	T	G	G	A	G	C	[illegible]	A	[illegible]	C	A	C	T	C	T	T	A	T	A	A	T	T	G	C	C
Lemur catta	A	A	G	C	T	T	C	A	T	A	G	G	A	G	C	[illegible]	A	[illegible]	C	A	T	T	C	T	A	A	T	A	A	T	C	G	C	A



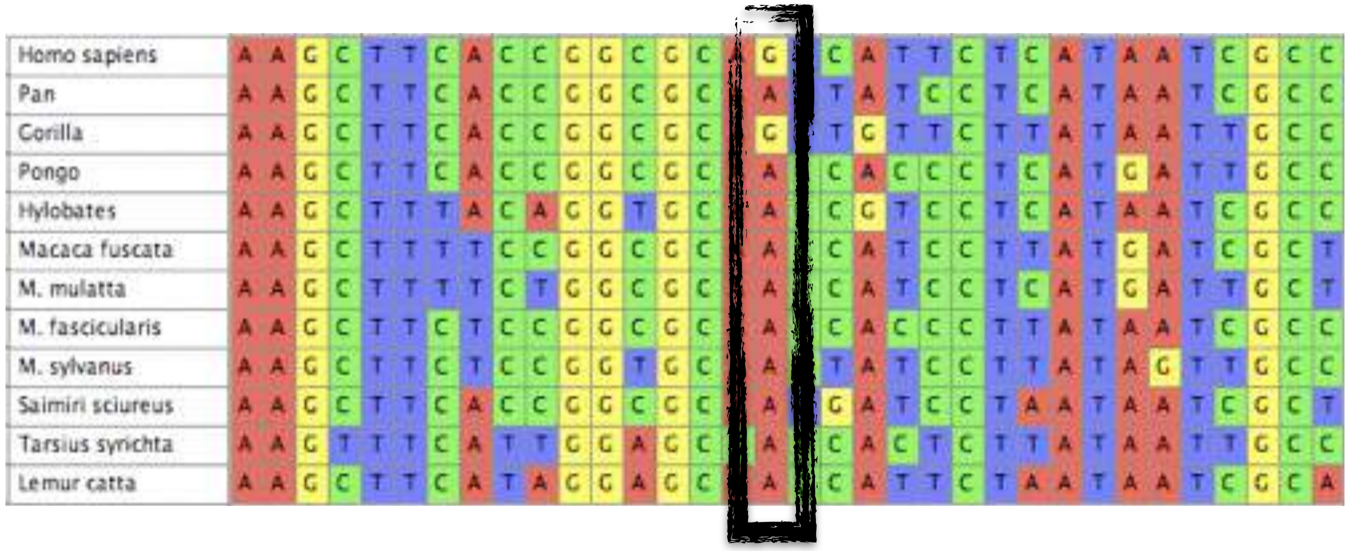
Markov process on a tree

Definition

Let $\mathcal{T} = (T;\phi)$ be a phylogenetic X-tree with root ρ. Let π^ρ be a distribution over $C = \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$ and, for each $e \in E(T)$ (away from the root), let $P(e) = [P(e)_{\alpha,\beta}]$ be a Markov transition matrix over C. Let $\theta = (\pi^\rho; P(e), e \in E(T))$. The distribution of a state vector $\chi : X \to C$ at the leaves is defined as

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

A natural choice is $P(e) = e^{\mu_e t_e Q}$ for a fixed rate matrix Q.



Homo sapiens	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	C	A	T	T	C	T	C	A	T	A	A	T	C	G	C	C
Pan	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	T	A	T	C	C	T	C	A	T	A	A	T	C	G	C	C
Gorilla	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	G	T	T	G	T	T	C	T	T	A	T	A	A	T	T	G	C	C
Pongo	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	C	A	T	G	A	T	T	G	C	C
Hylobates	A	A	G	C	T	T	T	A	C	A	G	G	T	G	C	A	A	C	C	G	T	C	C	T	C	A	T	A	A	T	C	G	C	C
Macaca fuscata	A	A	G	C	T	T	T	T	C	C	G	G	C	G	C	A	A	C	C	A	T	C	C	T	T	A	T	G	A	T	C	G	C	T
M. mulatta	A	A	G	C	T	T	T	T	C	T	G	G	C	G	C	A	A	C	C	A	T	C	C	T	C	A	T	G	A	T	T	G	C	T
M. fascicularis	A	A	G	C	T	T	C	T	C	C	G	G	C	G	C	A	A	C	C	A	C	C	C	T	T	A	T	A	A	T	C	G	C	C
M. sylvanus	A	A	G	C	T	T	C	T	C	C	G	G	T	G	C	A	A	C	T	A	T	C	C	T	T	A	T	A	G	T	T	G	C	C
Saimiri sciureus	A	A	G	C	T	T	C	A	C	C	G	G	C	G	C	A	A	T	G	A	T	C	C	T	A	A	T	A	A	T	C	G	C	T
Tarsius syrichta	A	A	G	T	T	T	C	A	T	T	G	G	A	G	C	C	A	C	C	A	C	T	C	T	T	A	T	A	A	T	T	G	C	C
Lemur catta	A	A	G	C	T	T	C	A	T	A	G	G	A	G	C	A	A	C	C	A	T	T	C	T	A	A	T	A	A	T	C	G	C	A

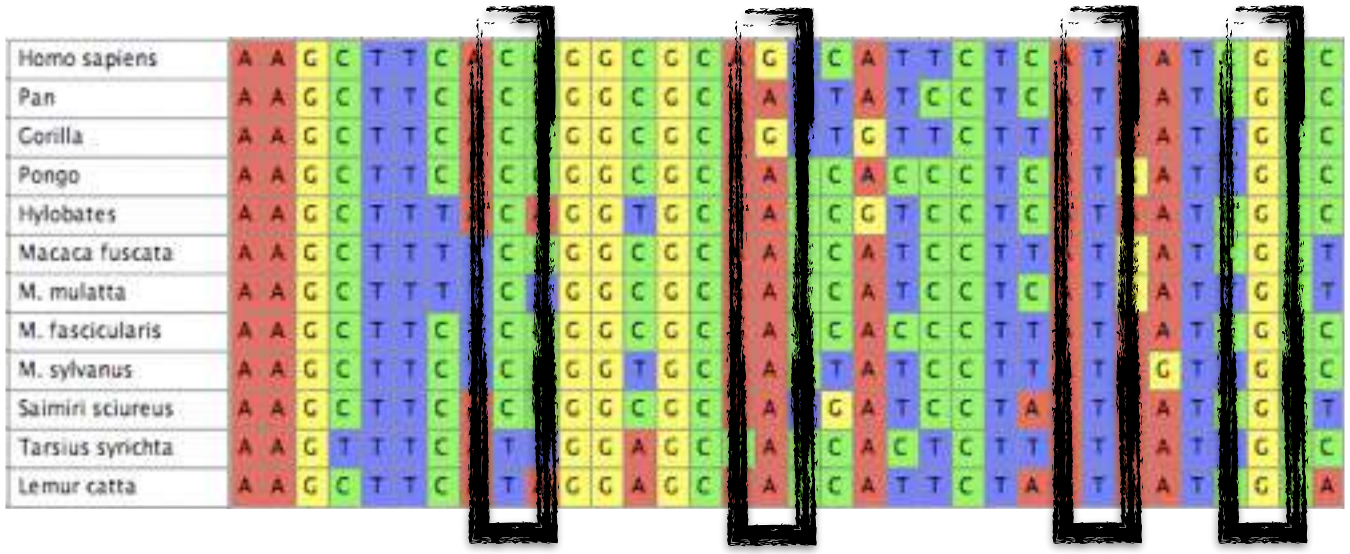
Markov process on a tree

Definition

Let $\mathcal{T} = (T; \phi)$ be a phylogenetic X-tree with root ρ. Let π^{ρ} be a distribution over $C = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ and, for each $e \in E(T)$ (away from the root), let $P(e) = [P(e)_{\alpha,\beta}]$ be a Markov transition matrix over C. Let $\theta = (\pi^{\rho}; P(e), e \in E(T))$. The distribution of a state vector $\chi : X \to C$ at the leaves is defined as

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

A natural choice is $P(e) = e^{\mu_e t_e Q}$ for a fixed rate matrix Q.



Homo sapiens	A	A	G	C	T	T	C	A	C	[illegible]	G	G	C	G	C	A	G	[illegible]	C	A	T	T	C	T	C	A	T	[illegible]	A	T	[illegible]	G	[illegible]	C
Pan	A	A	G	C	T	T	C	[illegible]	C	[illegible]	G	G	C	G	C	[illegible]	A	[illegible]	T	A	T	C	C	T	C	[illegible]	T	[illegible]	A	T	[illegible]	G	[illegible]	C
Gorilla	A	A	G	C	T	T	C	[illegible]	C	[illegible]	G	G	C	G	C	[illegible]	G	[illegible]	T	G	T	T	C	T	T	[illegible]	T	[illegible]	A	T	[illegible]	G	[illegible]	C
Pongo	A	A	G	C	T	T	C	[illegible]	C	[illegible]	G	G	C	G	C	[illegible]	A	[illegible]	C	A	C	C	C	T	C	[illegible]	T	[illegible]	A	T	[illegible]	G	[illegible]	C
Hylobates	A	A	G	C	T	T	T	[illegible]	C	A	G	G	T	G	C	[illegible]	A	[illegible]	C	G	T	C	C	T	C	[illegible]	T	[illegible]	A	T	[illegible]	G	[illegible]	C
Macaca fuscata	A	A	G	C	T	T	T	[illegible]	C	[illegible]	G	G	C	G	C	[illegible]	A	[illegible]	C	A	T	C	C	T	T	[illegible]	T	[illegible]	A	T	[illegible]	G	[illegible]	T
M. mulatta	A	A	G	C	T	T	T	[illegible]	C	[illegible]	G	G	C	G	C	[illegible]	A	[illegible]	C	A	T	C	C	T	C	[illegible]	T	[illegible]	A	T	[illegible]	G	[illegible]	T
M. fascicularis	A	A	G	C	T	T	C	[illegible]	C	[illegible]	G	G	C	G	C	[illegible]	A	[illegible]	C	A	C	C	C	T	T	[illegible]	T	[illegible]	A	T	[illegible]	G	[illegible]	C
M. sylvanus	A	A	G	C	T	T	C	[illegible]	C	[illegible]	G	G	T	G	C	[illegible]	A	[illegible]	T	A	T	C	C	T	T	[illegible]	T	[illegible]	G	T	[illegible]	G	[illegible]	C
Saimiri sciureus	A	A	G	C	T	T	C	[illegible]	C	[illegible]	G	G	C	G	C	[illegible]	A	[illegible]	G	A	T	C	C	T	A	[illegible]	T	[illegible]	A	T	[illegible]	G	[illegible]	T
Tarsius syrichta	A	A	G	T	T	T	C	[illegible]	T	[illegible]	G	G	A	G	C	[illegible]	A	[illegible]	C	A	C	T	C	T	T	[illegible]	T	[illegible]	A	T	[illegible]	G	[illegible]	C
Lemur catta	A	A	G	C	T	T	C	[illegible]	T	[illegible]	G	G	A	G	C	[illegible]	A	[illegible]	C	A	T	T	C	T	A	[illegible]	T	[illegible]	A	T	[illegible]	G	[illegible]	A

k columns

=

k i.i.d. samples

Back to tree metrics

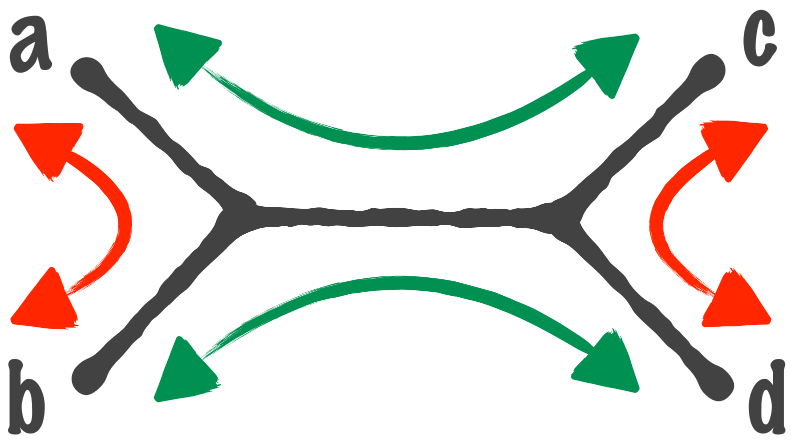
Definition

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

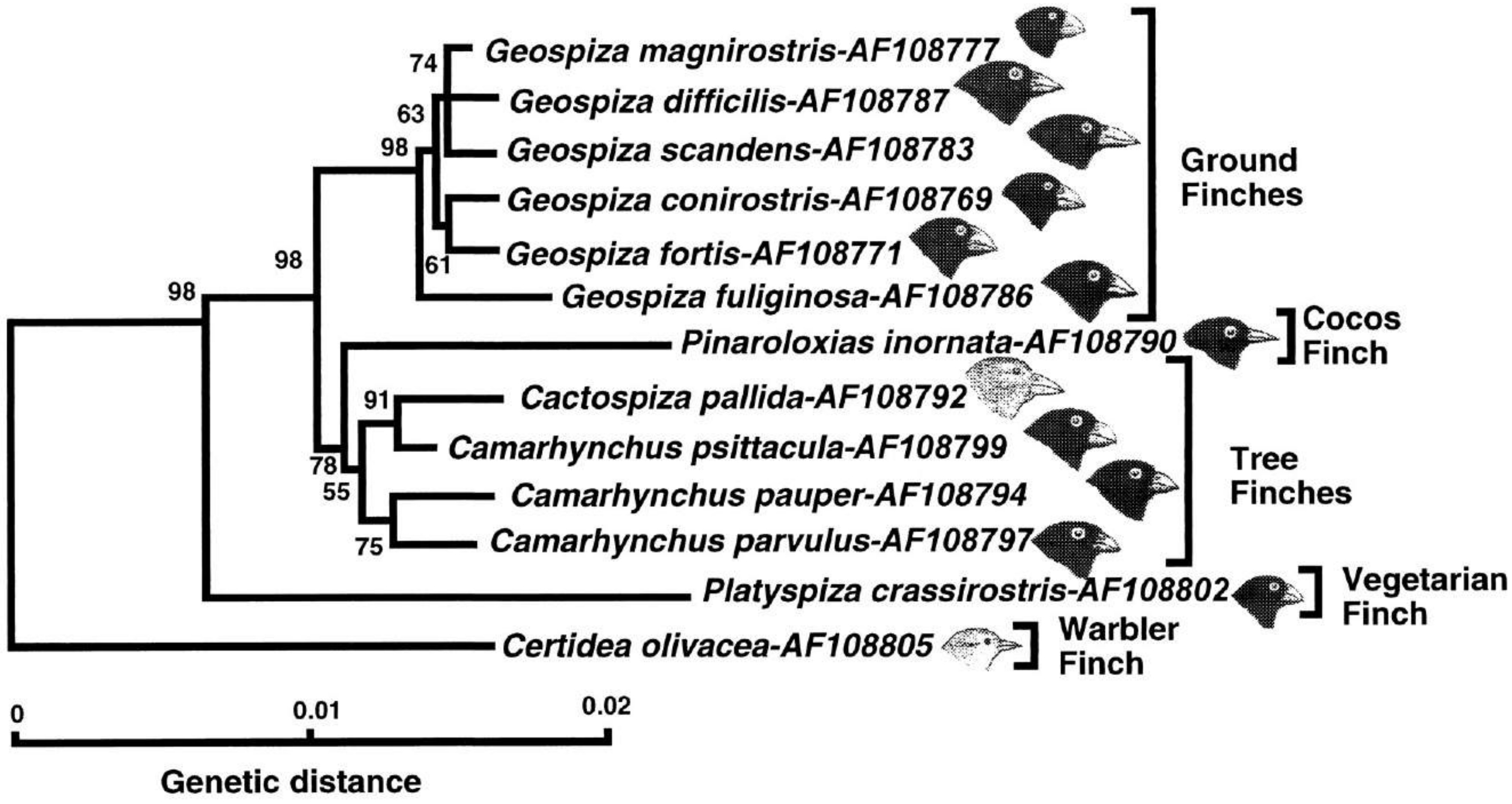
$$\delta(x, y) = -\log(\det(F^{xy})).$$

Theorem (Steel, AML (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 $\mathcal{T}$.

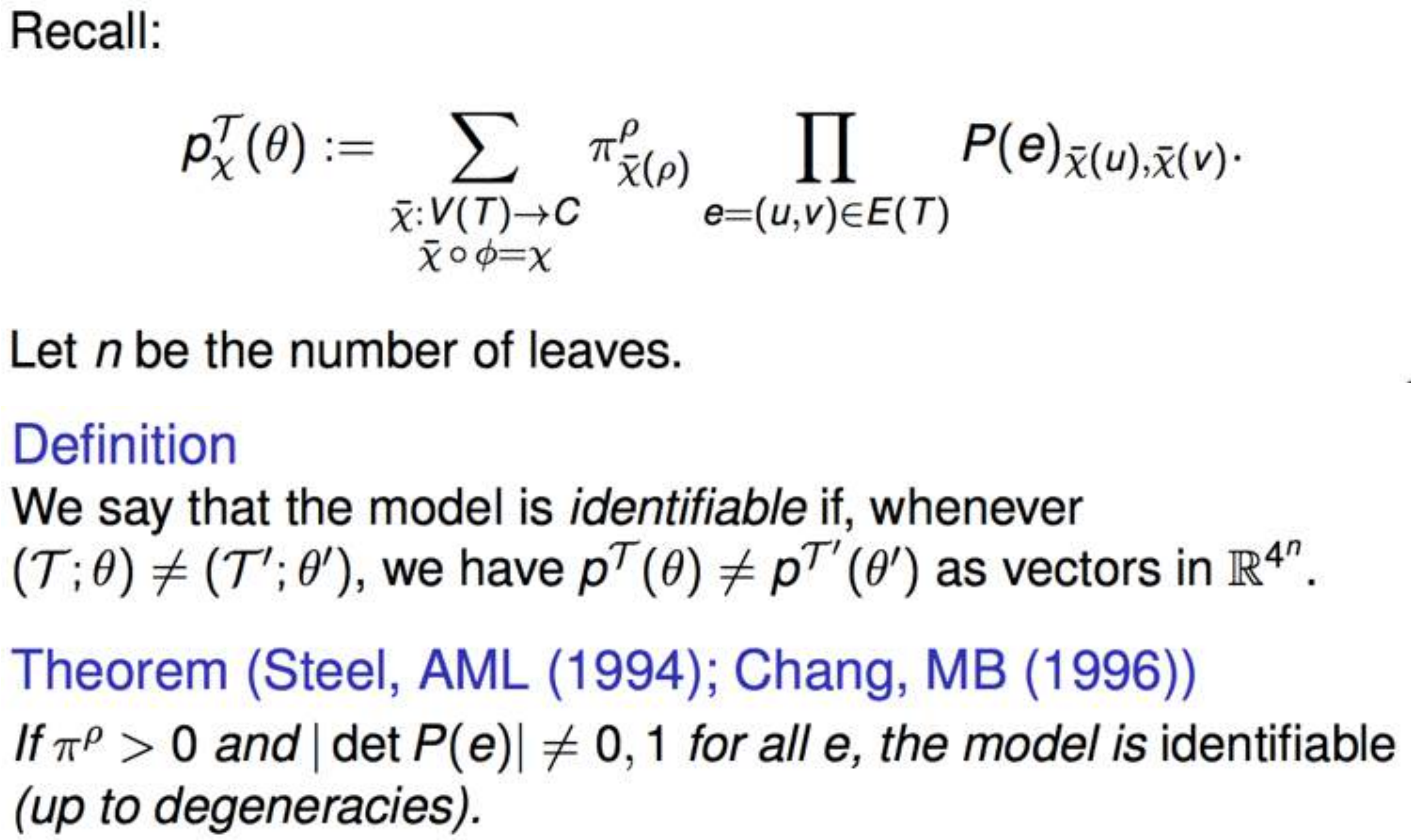


Back to Darwin’s finches


74
63
98
98
98
61
Geospiza magnirostris-AF108777
Geospiza difficilis-AF108787
Geospiza scandens-AF108783
Geospiza conirostris-AF108769
Geospiza fortis-AF108771
Geospiza fuliginosa-AF108786
Ground Finches
Pinaroloxias inornata-AF108790
Cocos Finch
91
78
55
75
Cactospiza pallida-AF108792
Camarhynchus psittacula-AF108799
Camarhynchus pauper-AF108794
Camarhynchus parvulus-AF108797
Tree Finches
Platyspiza crassirostris-AF108802
Vegetarian Finch
Certidea olivacea-AF108805
Warbler Finch
0
0.01
0.02
Genetic distance

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

Identifiability



Recall:

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

Let n be the number of leaves.

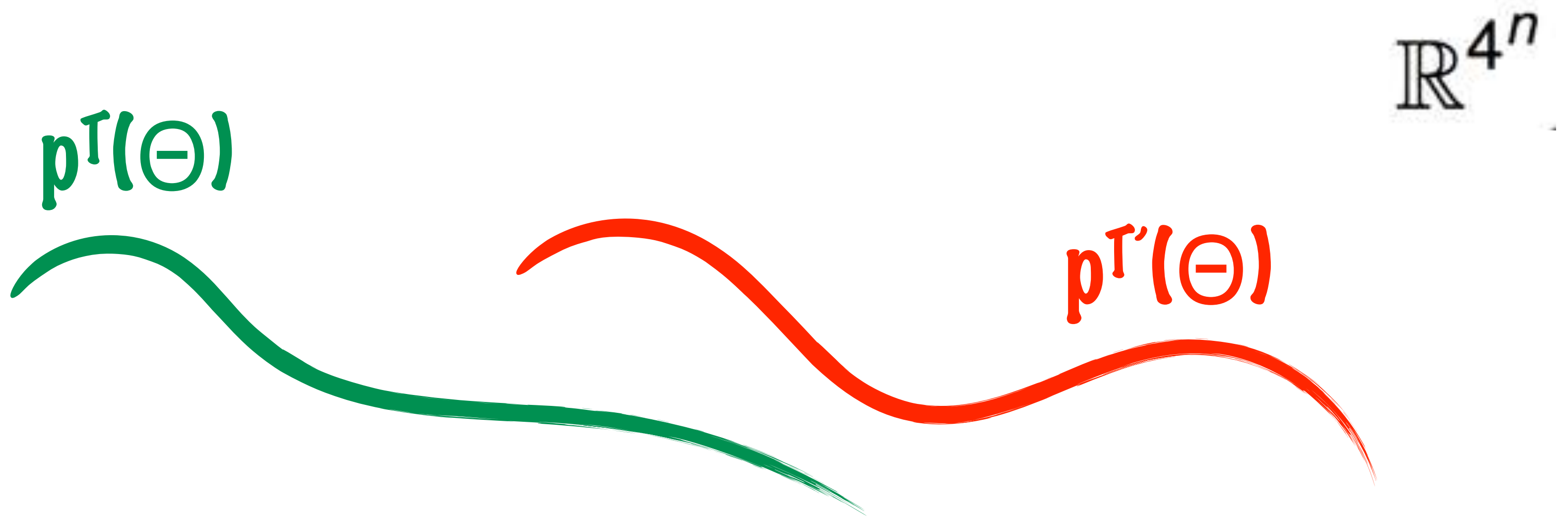
Definition

We say that the model is *identifiable* if, whenever $(\mathcal{T};\theta) \neq (\mathcal{T}';\theta')$, we have $p^{\mathcal{T}}(\theta) \neq p^{\mathcal{T}'}(\theta')$ as vectors in $\mathbb{R}^{4^n}$.

Theorem (Steel, AML (1994); Chang, MB (1996))

If $\pi^{\rho} > 0$ *and* $|\det P(e)| \neq 0, 1$ *for all e, the model is* identifiable *(up to degeneracies).*

Identifiability



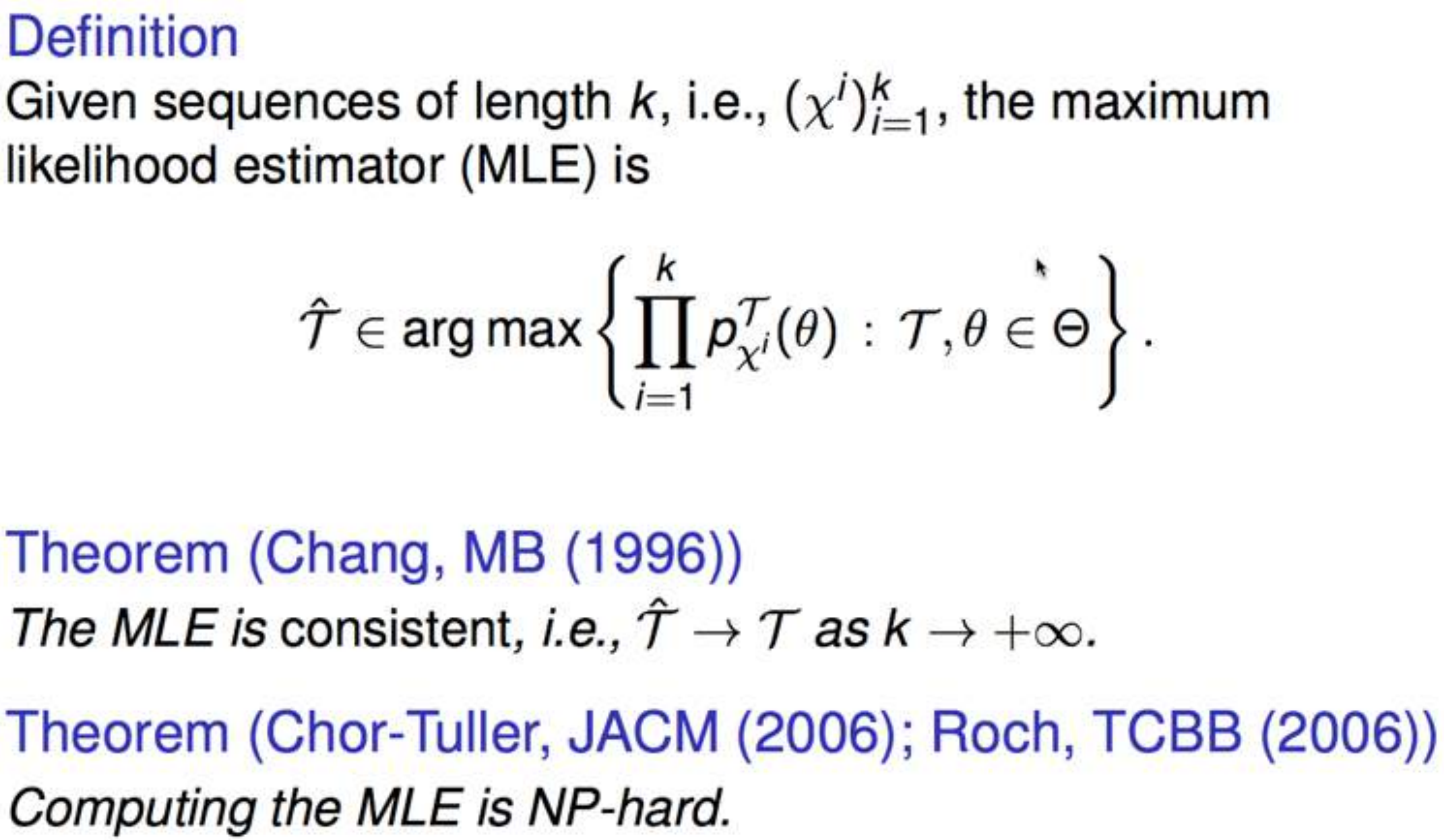
Definition

We say that the model is *identifiable* if, whenever $(\mathcal{T};\theta) \neq (\mathcal{T}';\theta')$, we have $p^{\mathcal{T}}(\theta) \neq p^{\mathcal{T}'}(\theta')$ as vectors in $\mathbb{R}^{4^n}$.

Theorem (Steel, AML (1994); Chang, MB (1996))

If $\pi^\rho > 0$ *and* $|\det P(e)| \neq 0, 1$ *for all e, the model is* identifiable *(up to degeneracies).*

Likelihood-based inference



Definition

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

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

Theorem (Chang, MB (1996))

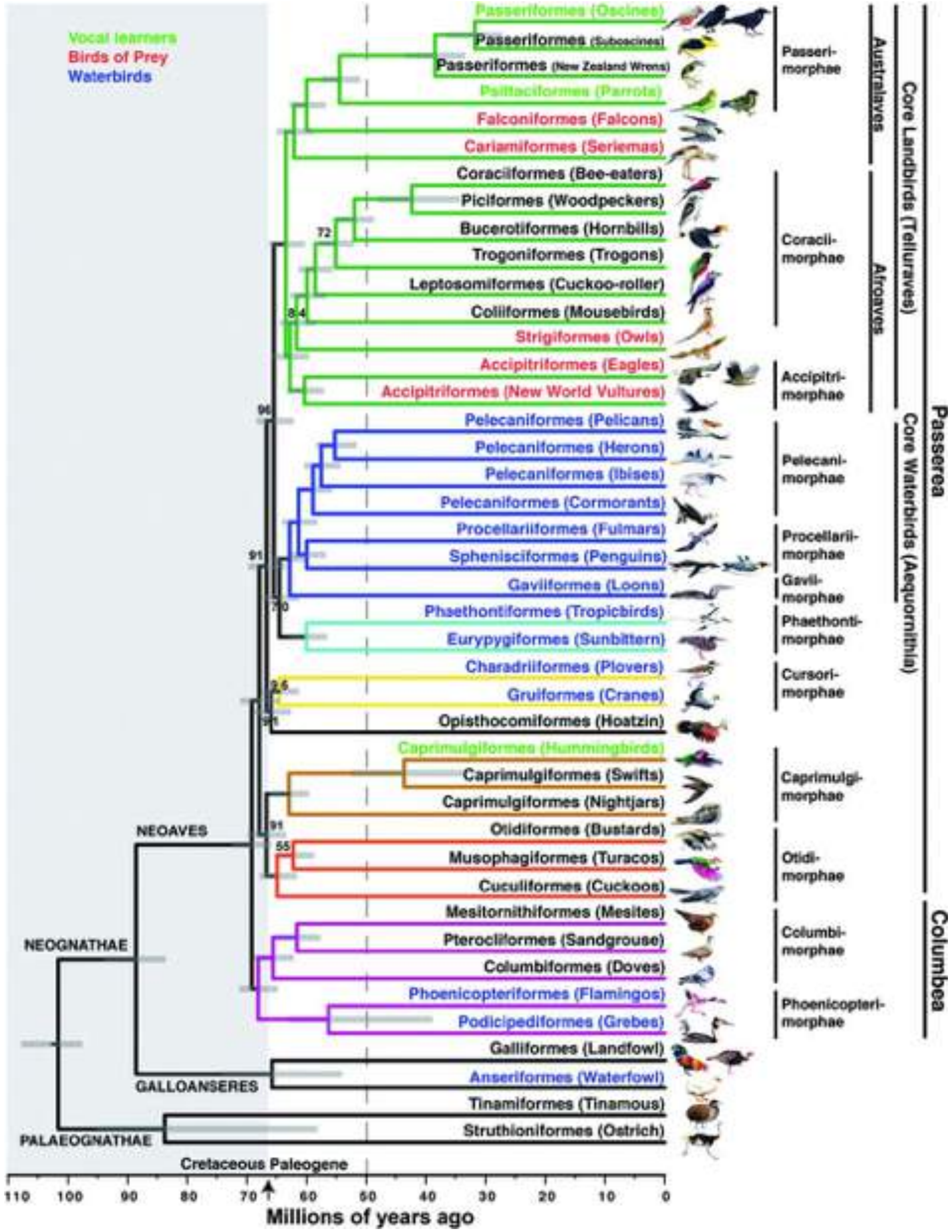
The MLE is consistent, *i.e.,* $\hat{\mathcal{T}} \to \mathcal{T}$ *as* $k \to +\infty$.

Theorem (Chor-Tuller, JACM (2006); Roch, TCBB (2006))

Computing the MLE is NP-hard.

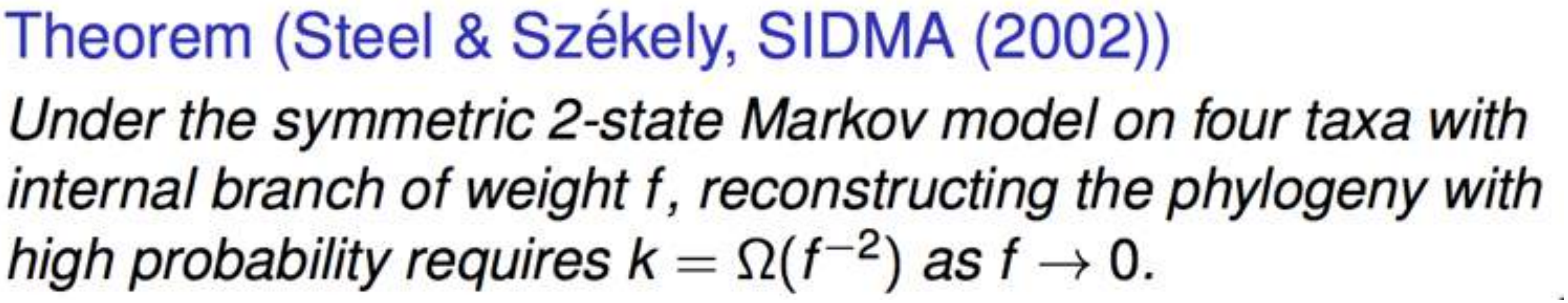
How much data do I need?

Adaptive radiation



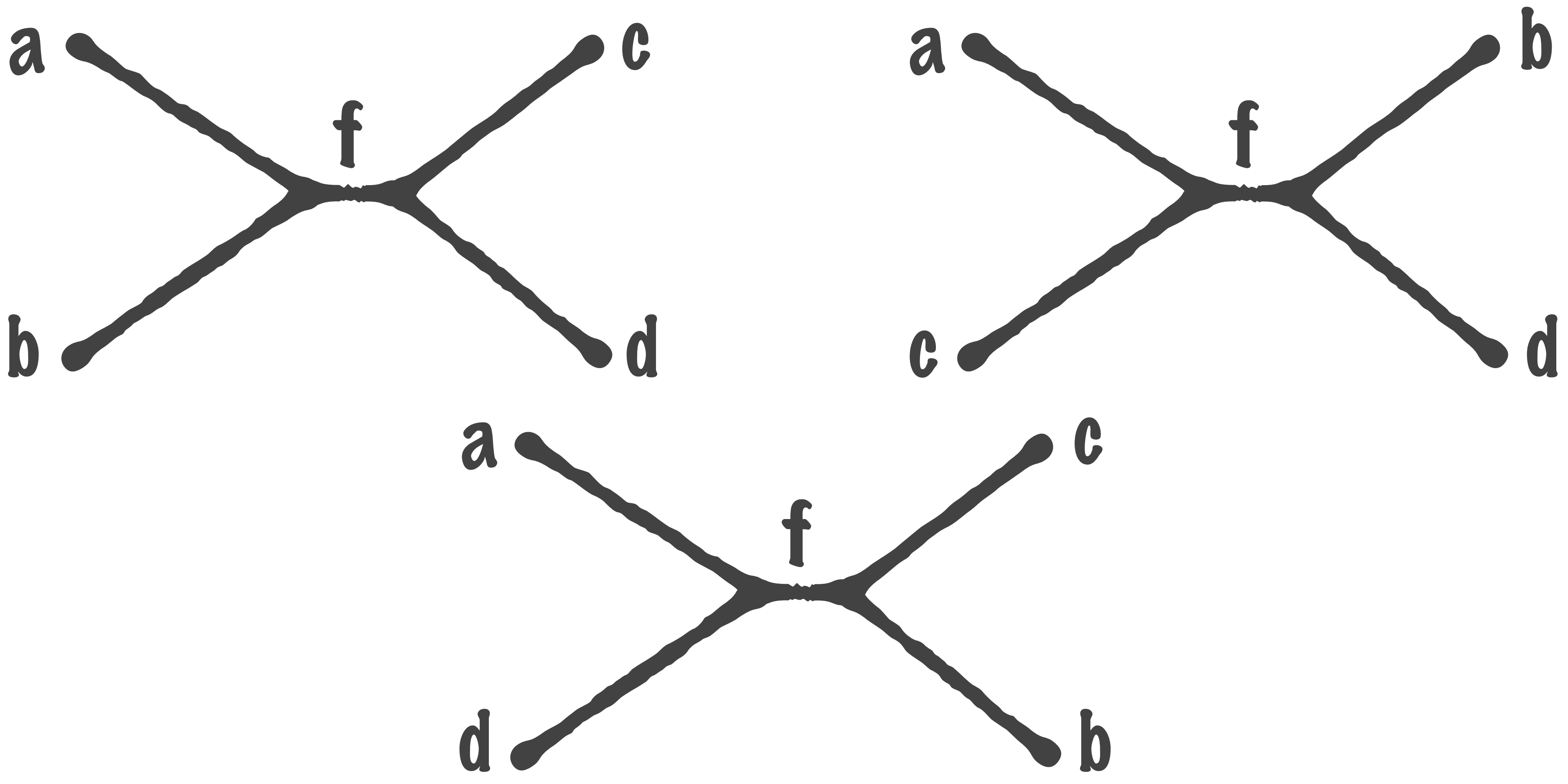
Genome-scale phylogeny of birds. (From: Erich D. Jarvis et al. Science 2014;346:1320-1331)

Short branches

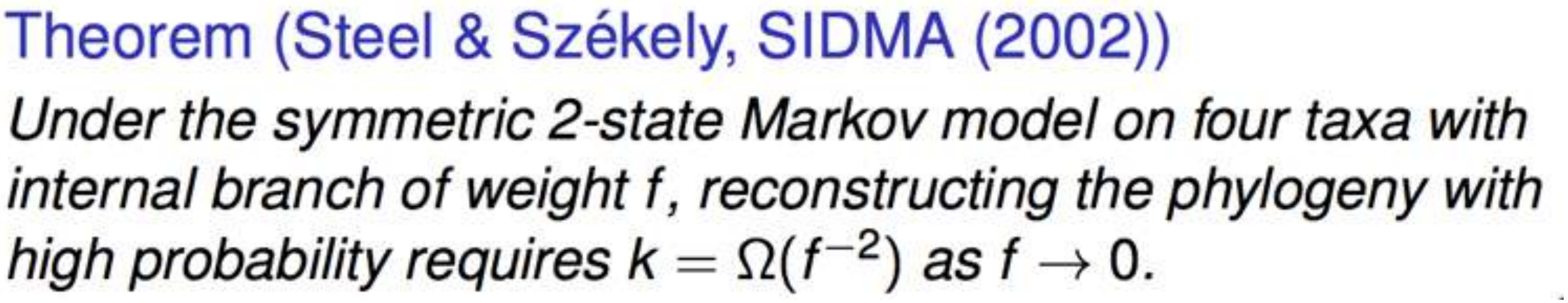


Theorem (Steel & Székely, SIDMA (2002))

Under the symmetric 2-state Markov model on four taxa with internal branch of weight f, reconstructing the phylogeny with high probability requires $k = \Omega(f^{-2})$ as $f \to 0$.

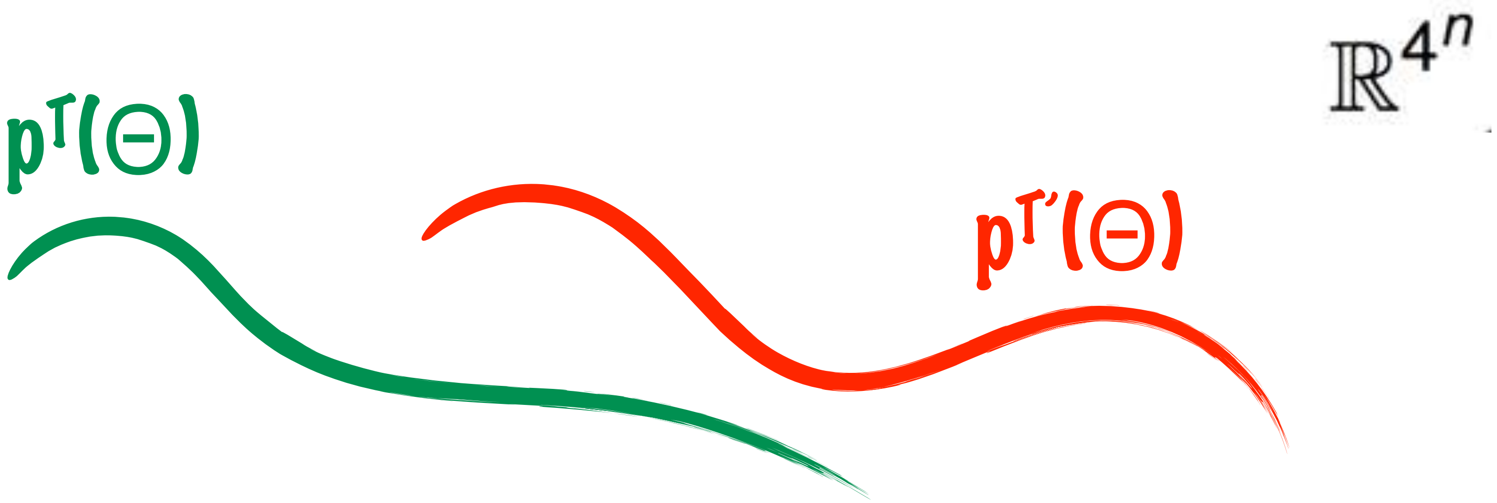


Short branches

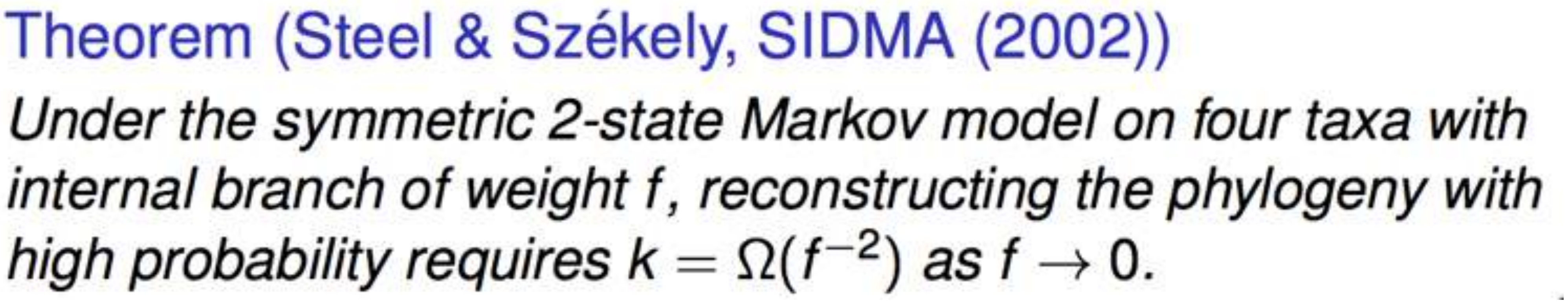


Theorem (Steel & Székely, SIDMA (2002))

Under the symmetric 2-state Markov model on four taxa with internal branch of weight f, reconstructing the phylogeny with high probability requires $k = \Omega(f^{-2})$ *as* $f \to 0$.

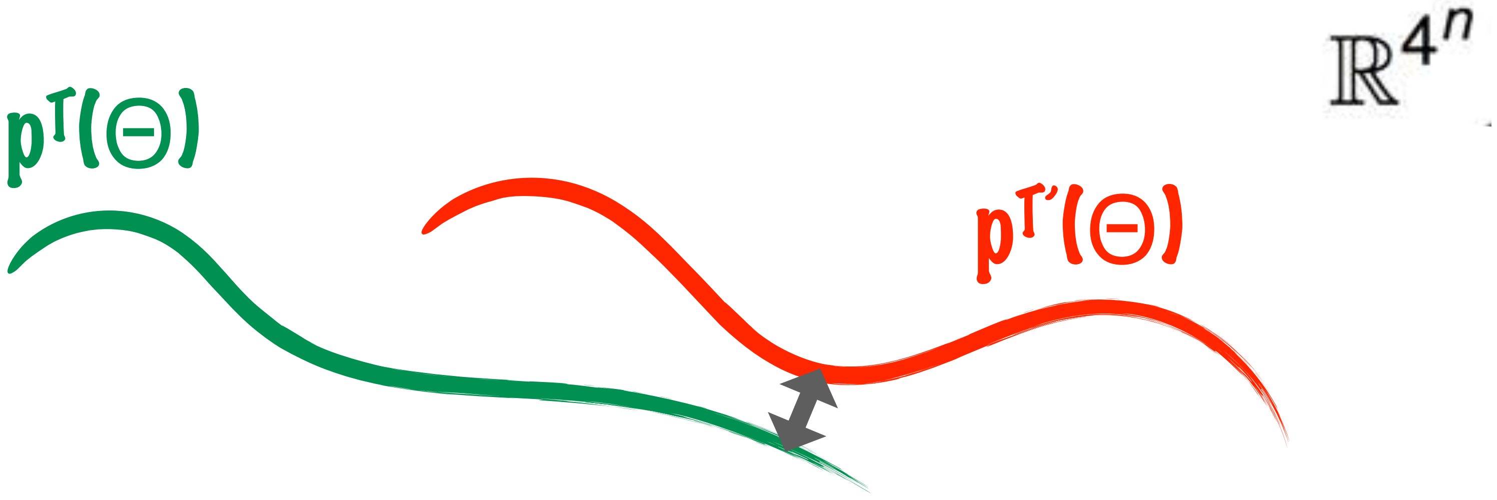


Short branches



Theorem (Steel & Székely, SIDMA (2002))

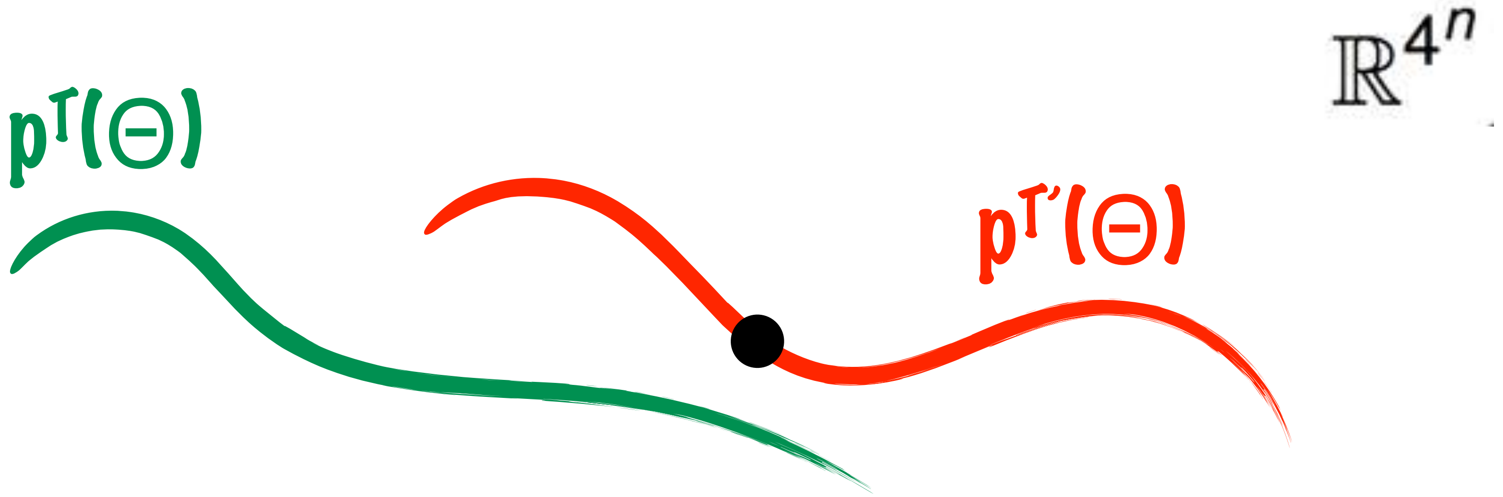
Under the symmetric 2-state Markov model on four taxa with internal branch of weight f, reconstructing the phylogeny with high probability requires $k = \Omega(f^{-2})$ *as* $f \to 0$.



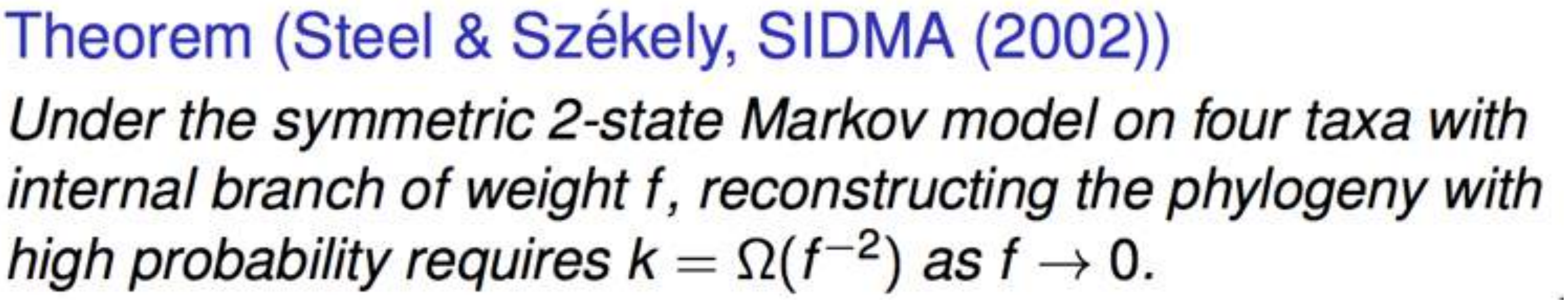
Short branches

Theorem (Steel & Székely, SIDMA (2002))

Under the symmetric 2-state Markov model on four taxa with internal branch of weight f, reconstructing the phylogeny with high probability requires $k = \Omega(f^{-2})$ *as* $f \to 0$.

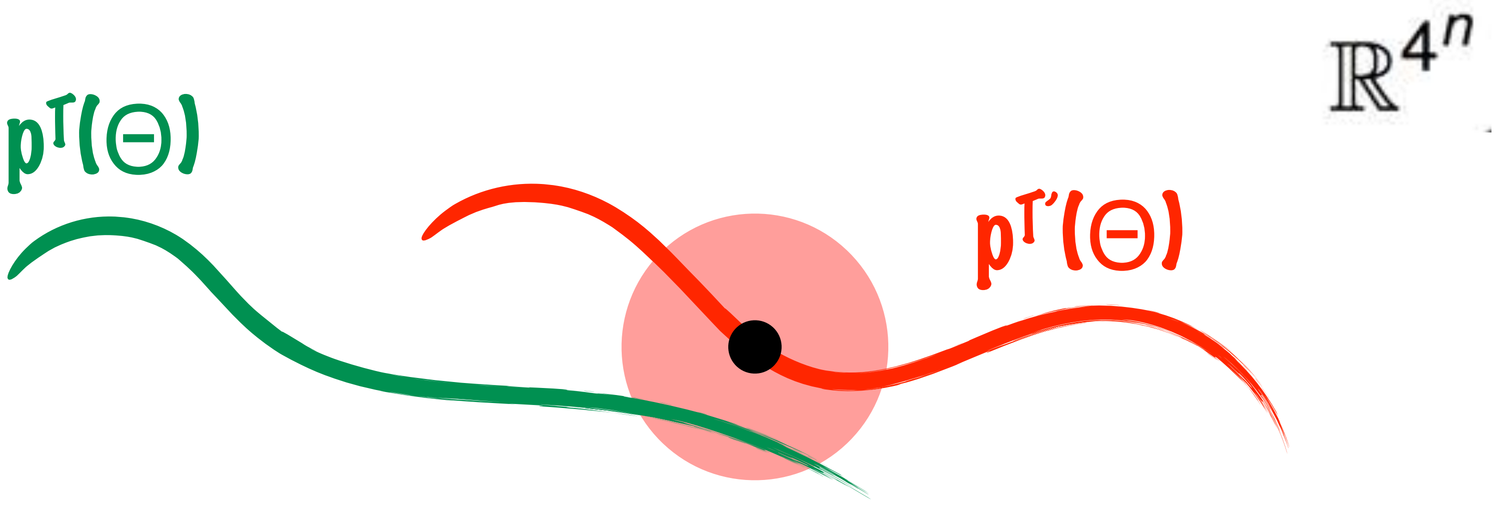


Short branches

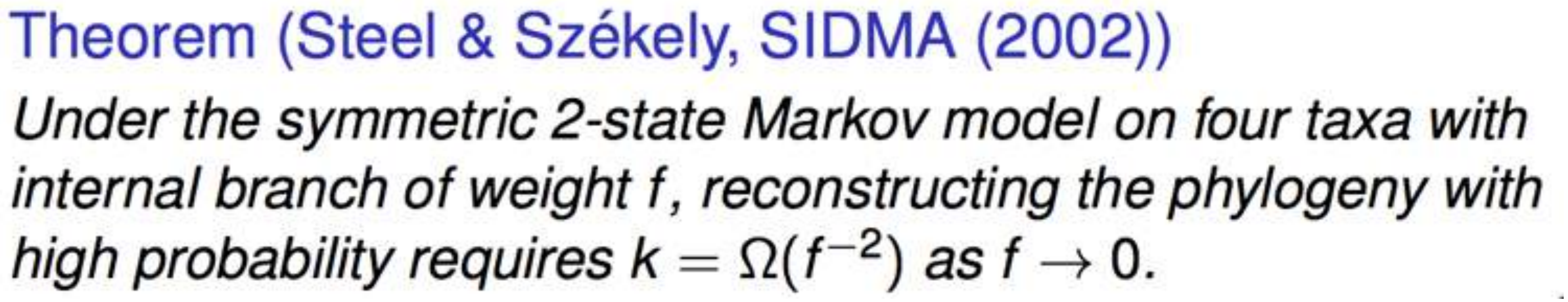


Theorem (Steel & Székely, SIDMA (2002))

Under the symmetric 2-state Markov model on four taxa with internal branch of weight f, reconstructing the phylogeny with high probability requires $k = \Omega(f^{-2})$ as $f \to 0$.

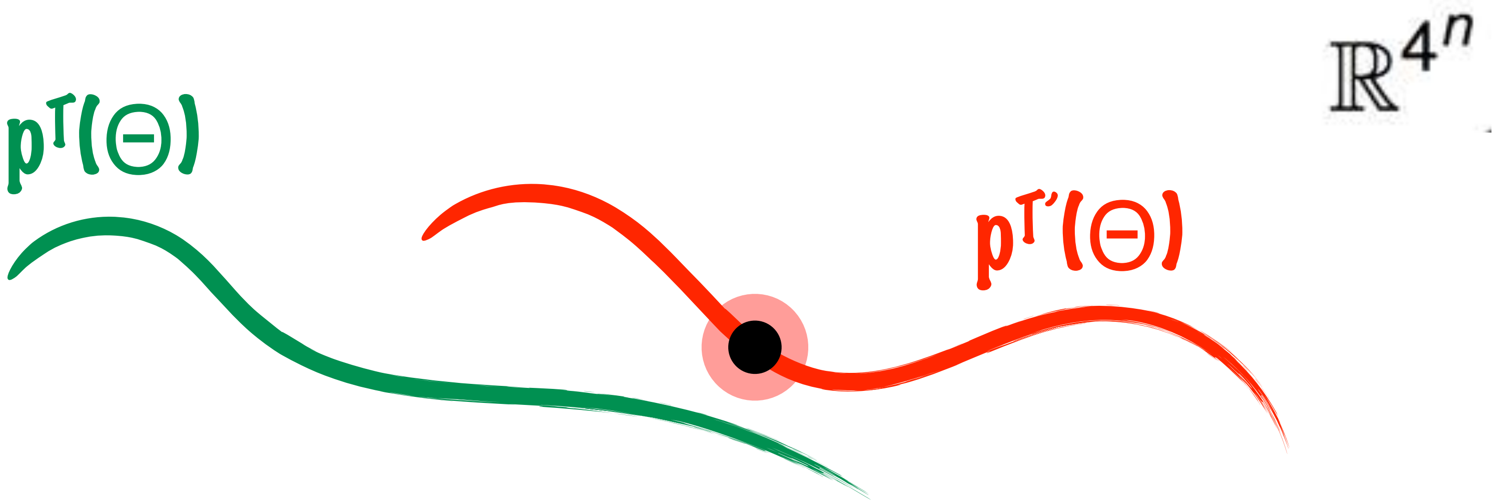


Short branches

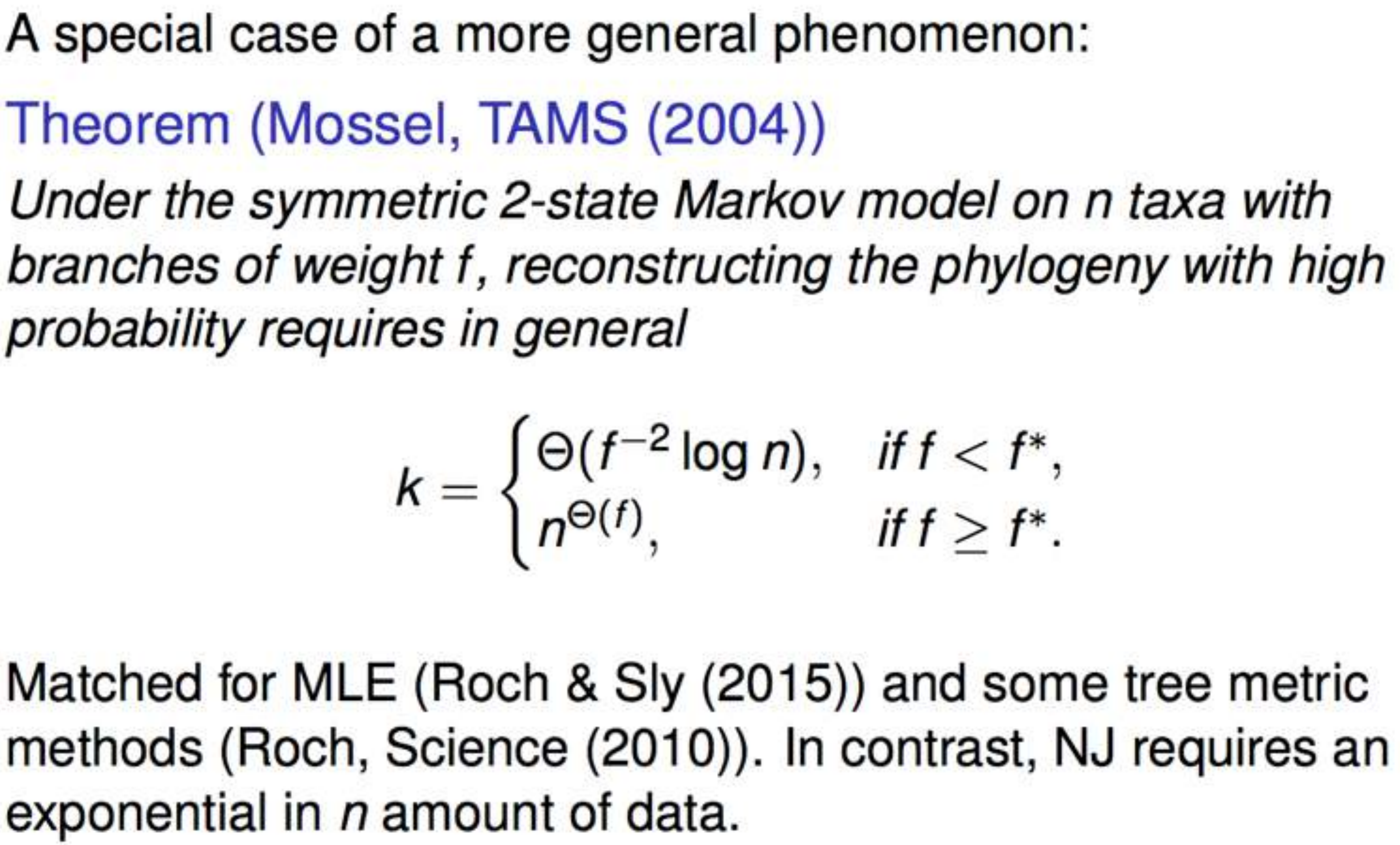


Theorem (Steel & Székely, SIDMA (2002))

Under the symmetric 2-state Markov model on four taxa with internal branch of weight f, reconstructing the phylogeny with high probability requires $k = \Omega(f^{-2})$ as $f \to 0$.



Depth



A special case of a more general phenomenon:

Theorem (Mossel, TAMS (2004))

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

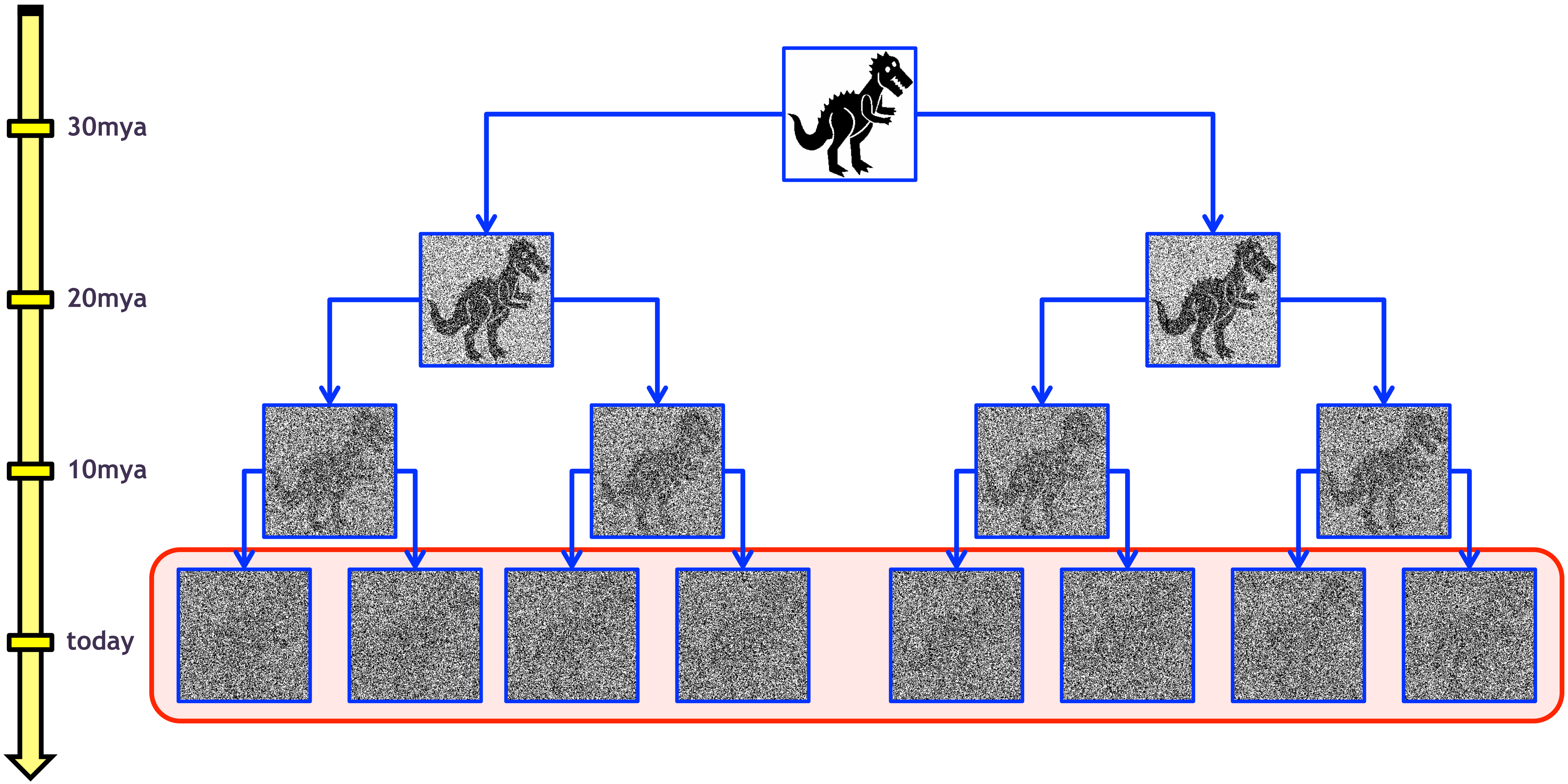
$$k = \begin{cases} \Theta(f^{-2}\log n), & \textit{if } f < f^*, \\ n^{\Theta(f)}, & \textit{if } f \geq f^*. \end{cases}$$

Matched for MLE (Roch & Sly (2015)) and some tree metric methods (Roch, Science (2010)). In contrast, NJ requires an exponential in *n* amount of data.

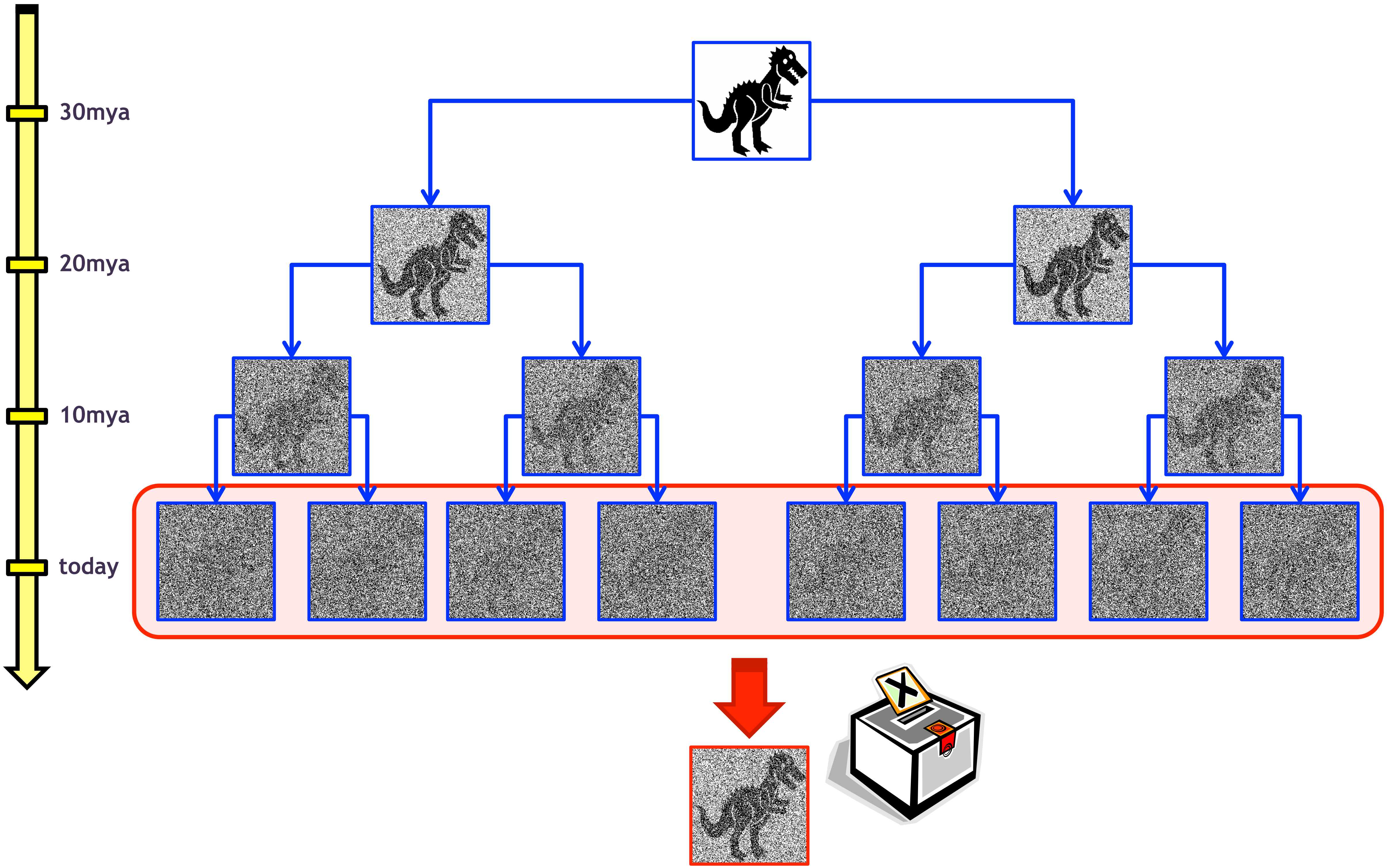
Correlation decay

Correlation decay

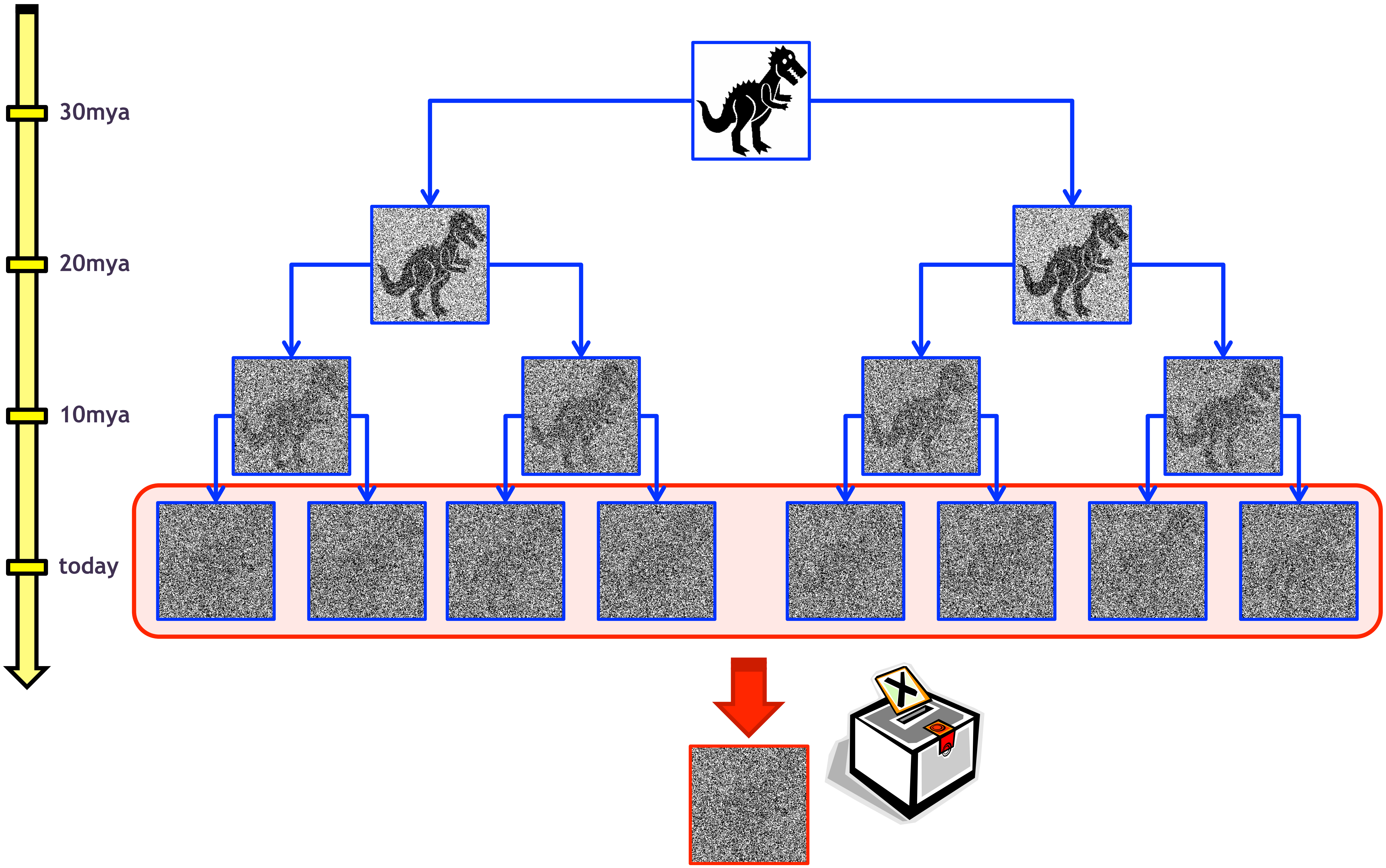
Correlation decay

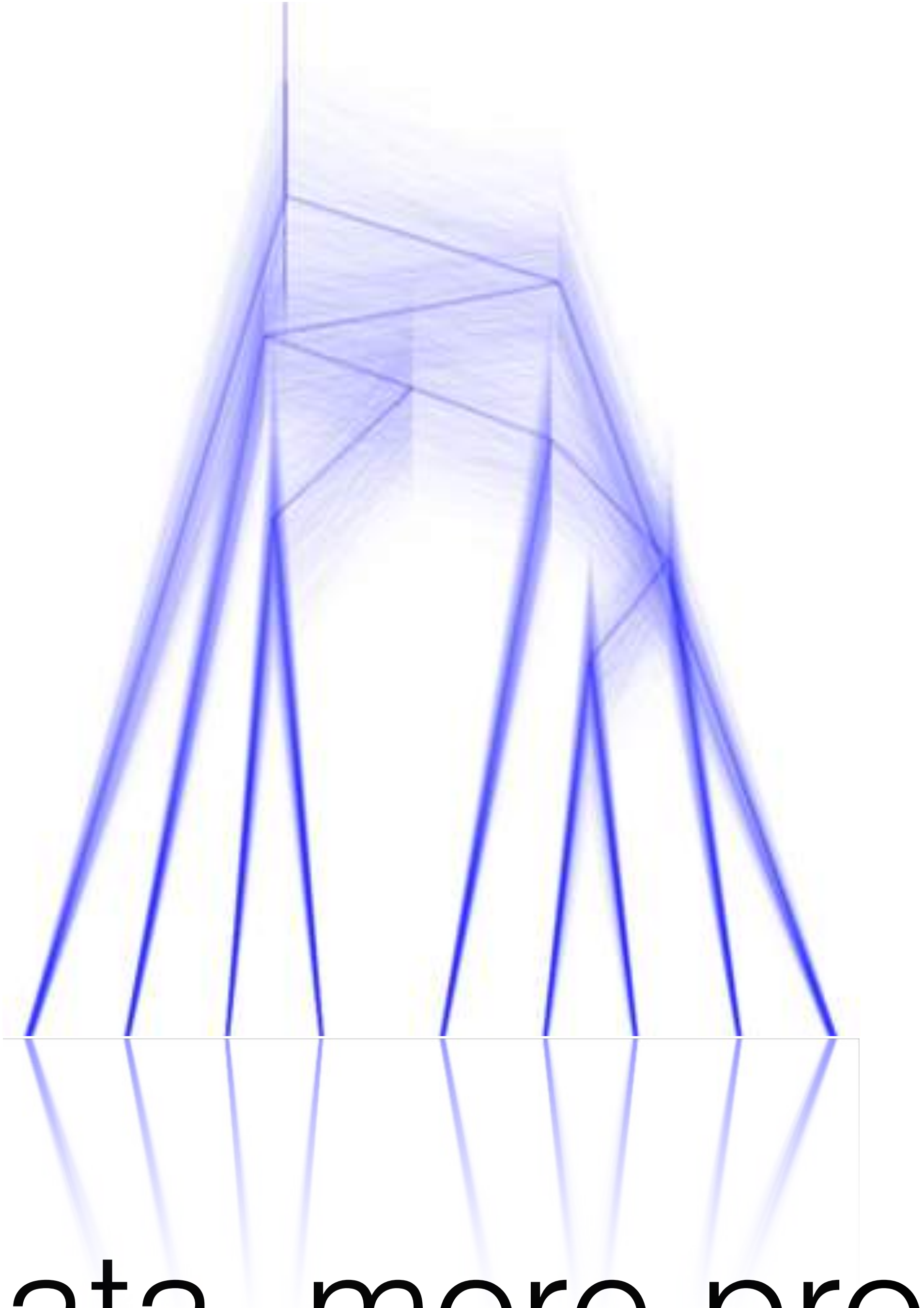


Correlation decay



Correlation decay

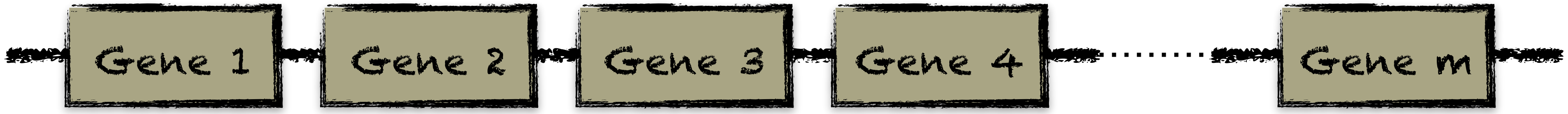




More data, more problems

Next-generation sequencing

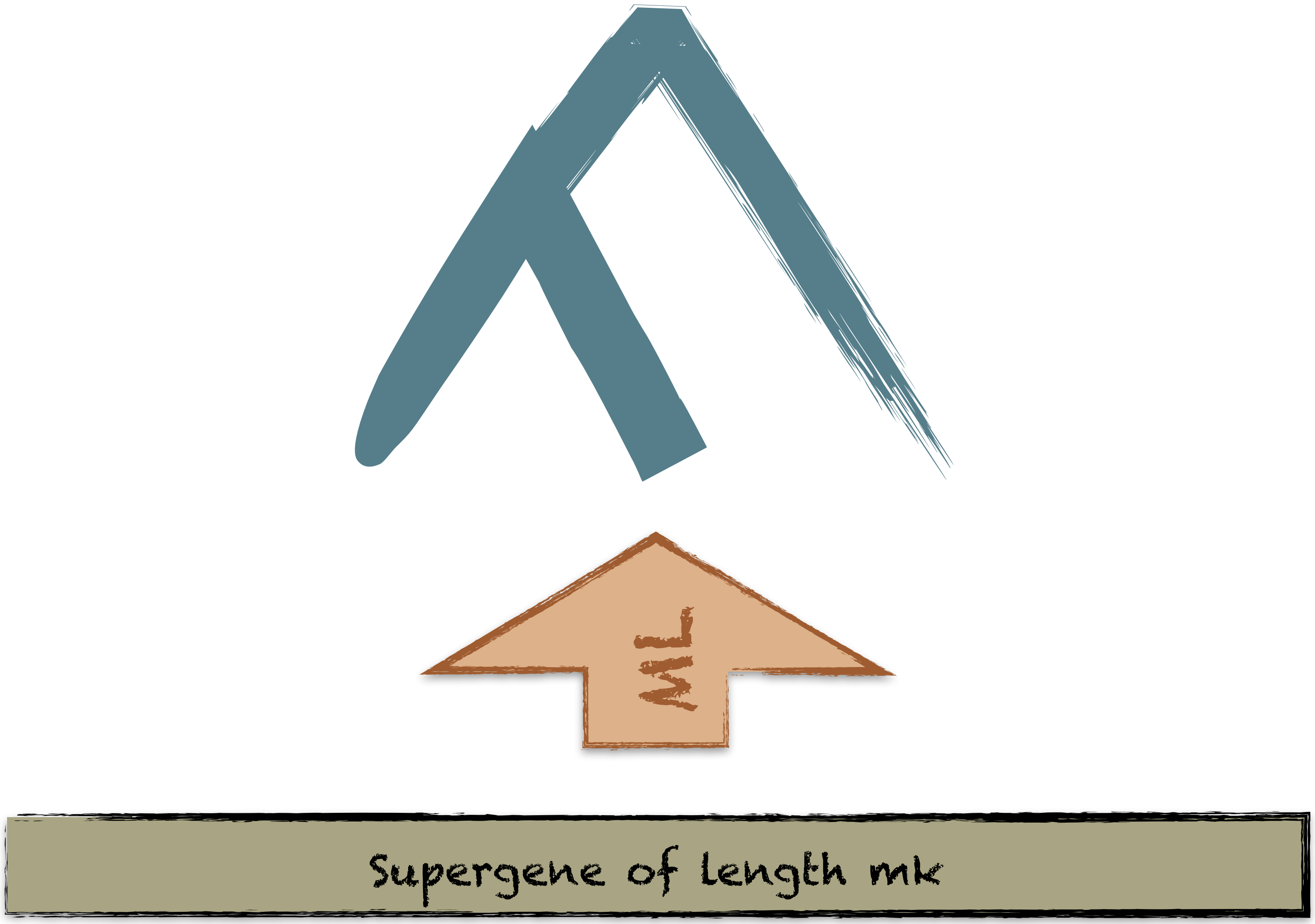
Concatenating genes



Concatenating genes

Supergene of length mk

Concatenating genes



ML

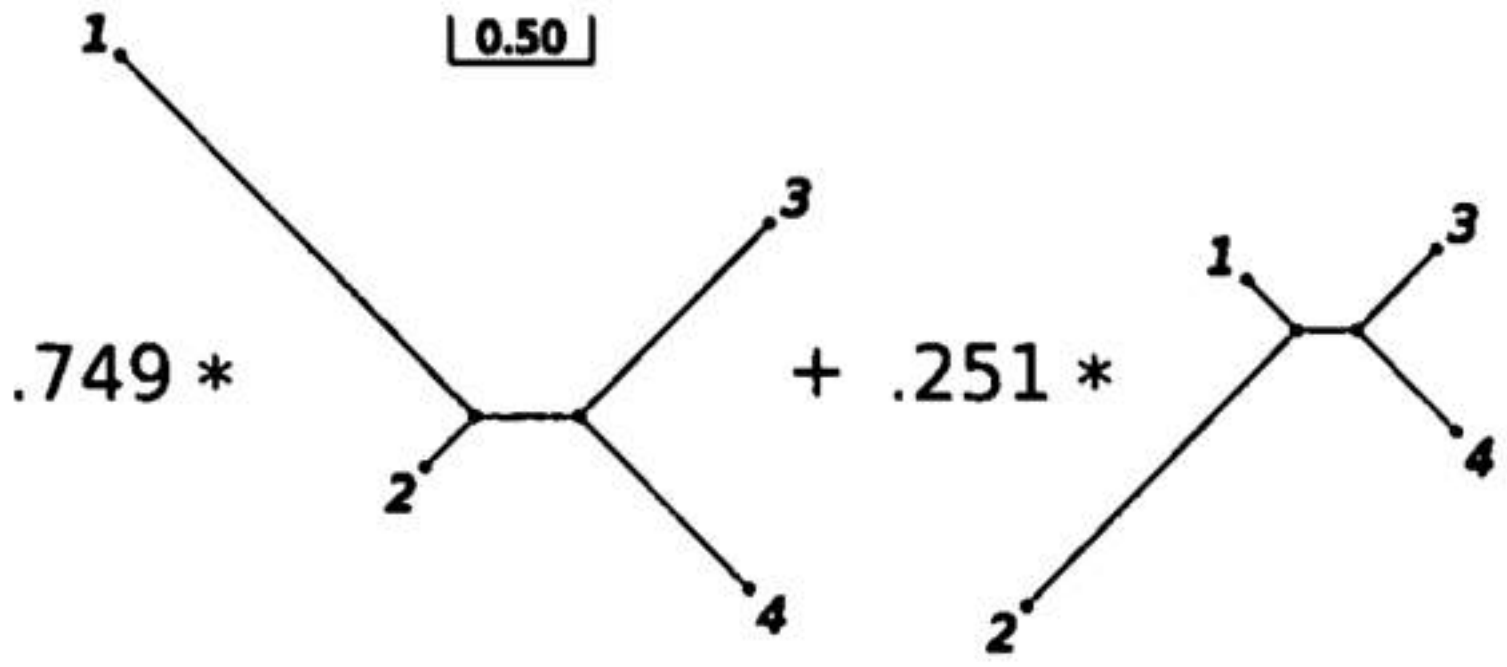
Supergene of length mk

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.

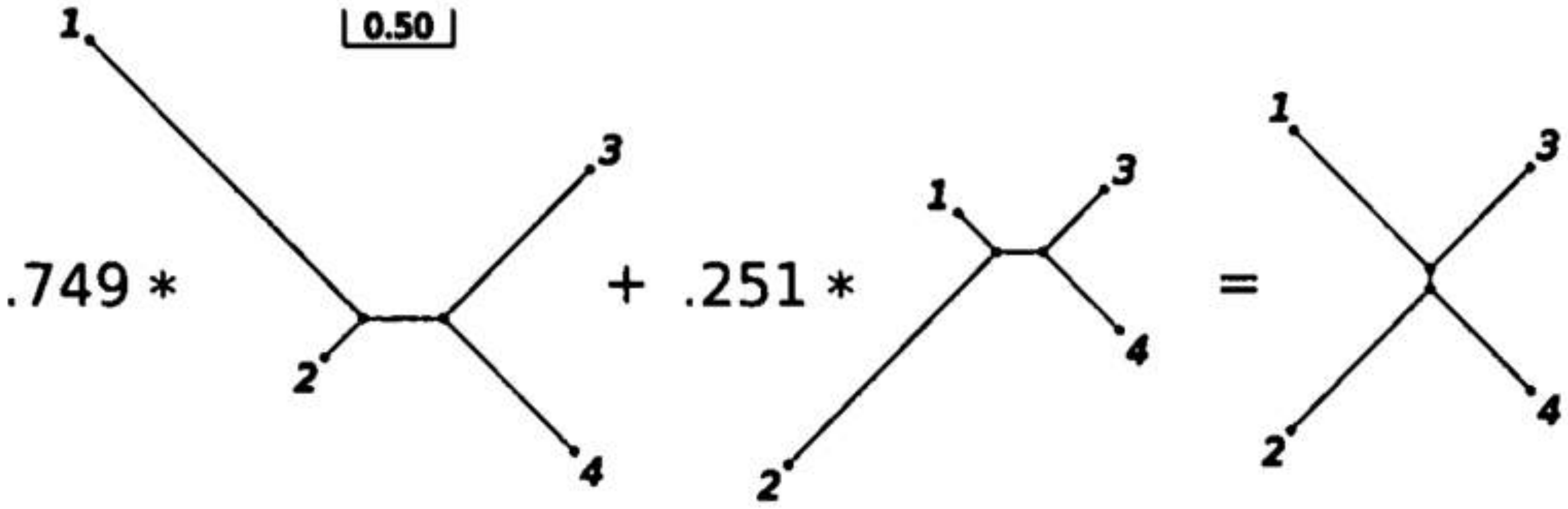


Mixed-up trees

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.

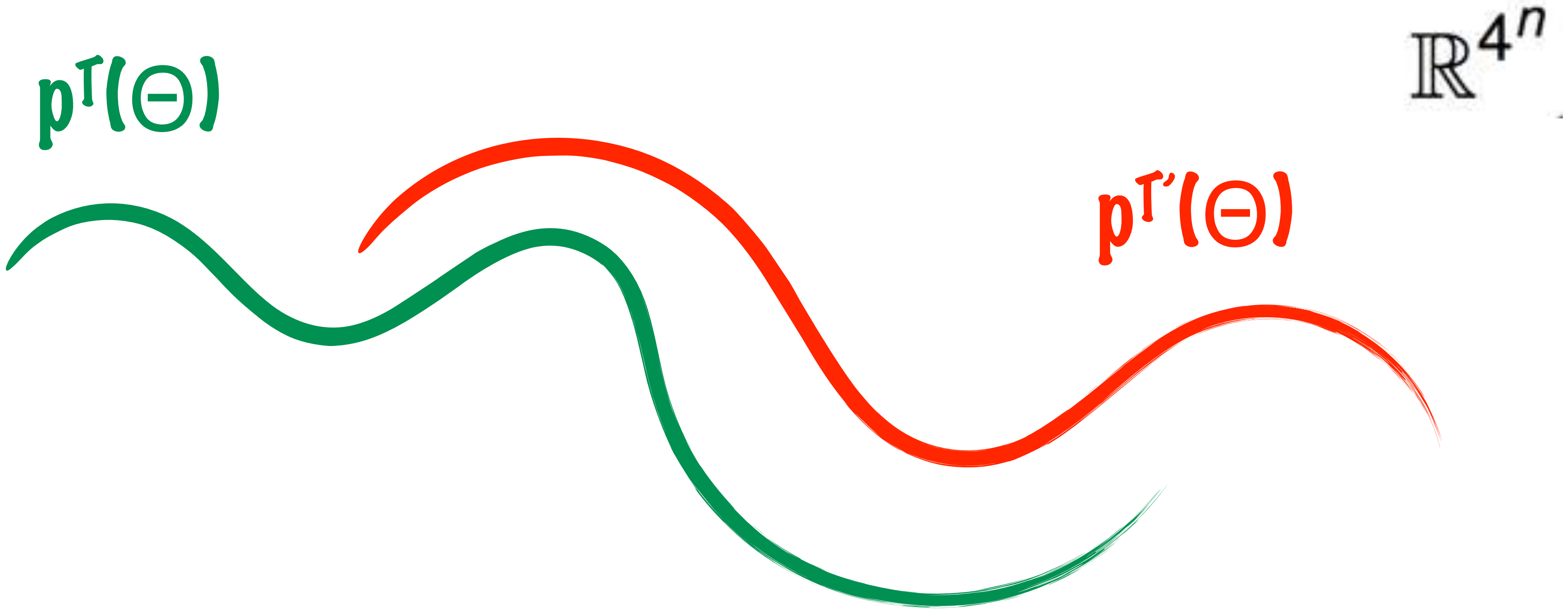


Mixed-up trees

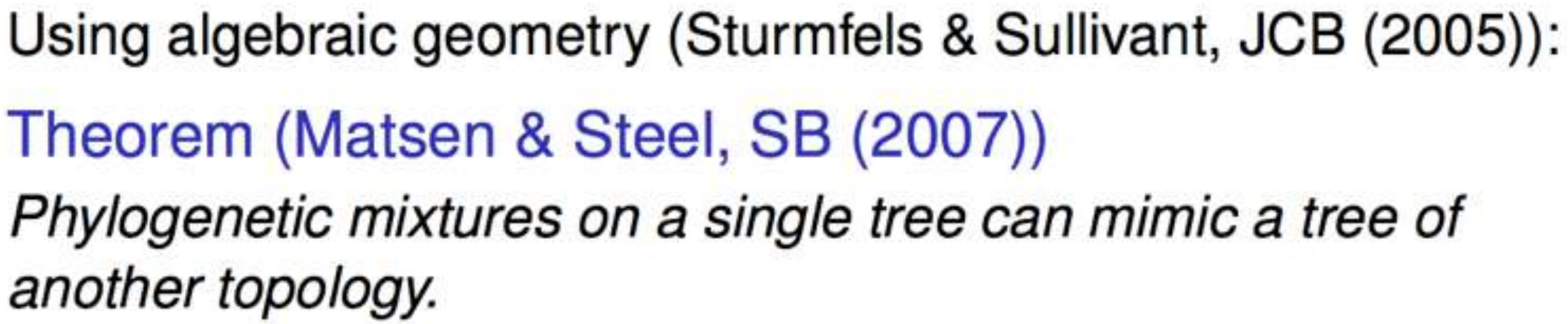
Using algebraic geometry (Sturmfels & Sullivant, JCB (2005)):

Theorem (Matsen & Steel, SB (2007))

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



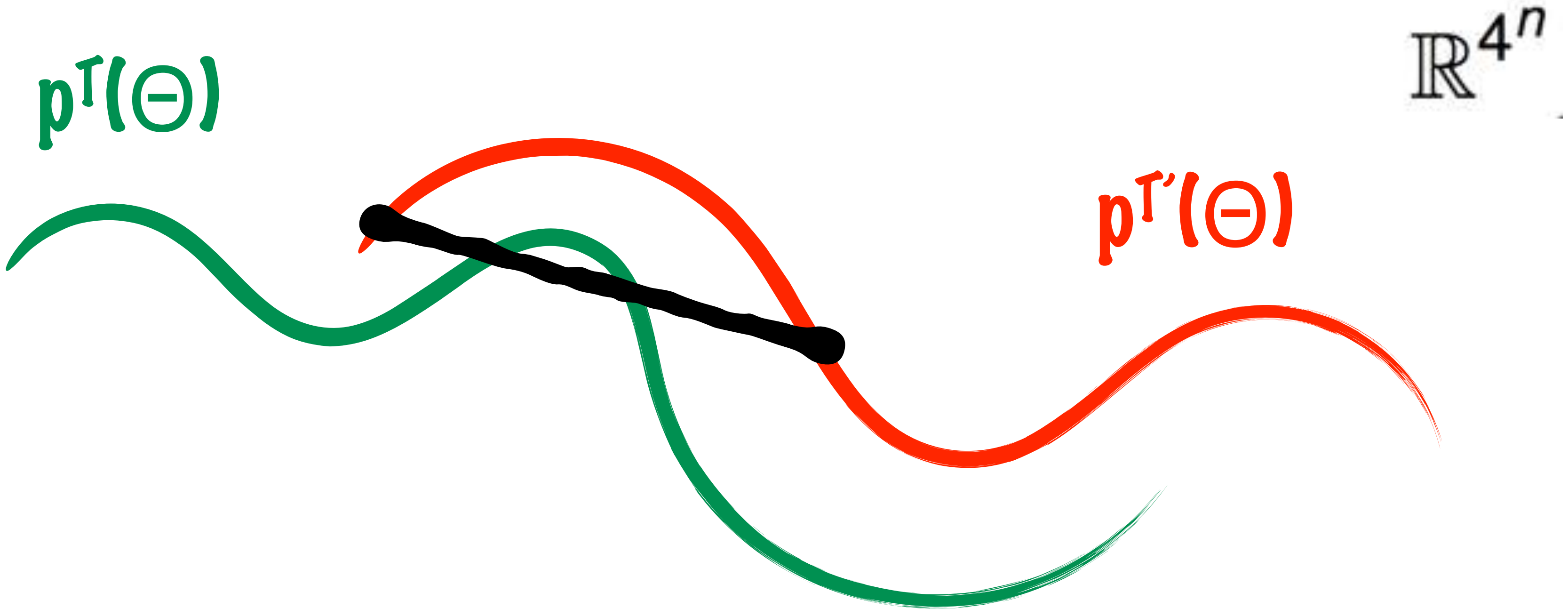
Mixed-up trees



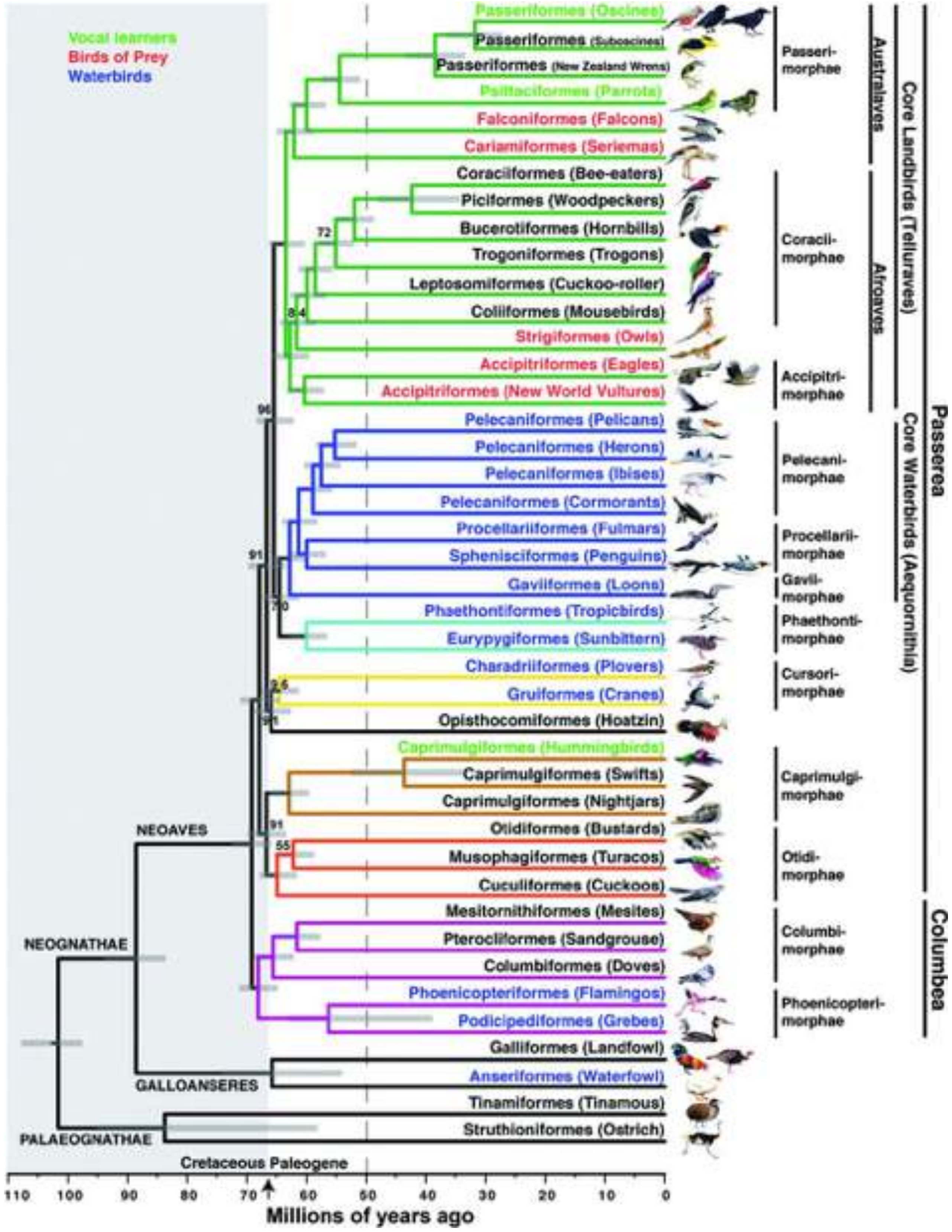
Using algebraic geometry (Sturmfels & Sullivant, JCB (2005)):

Theorem (Matsen & Steel, SB (2007))

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

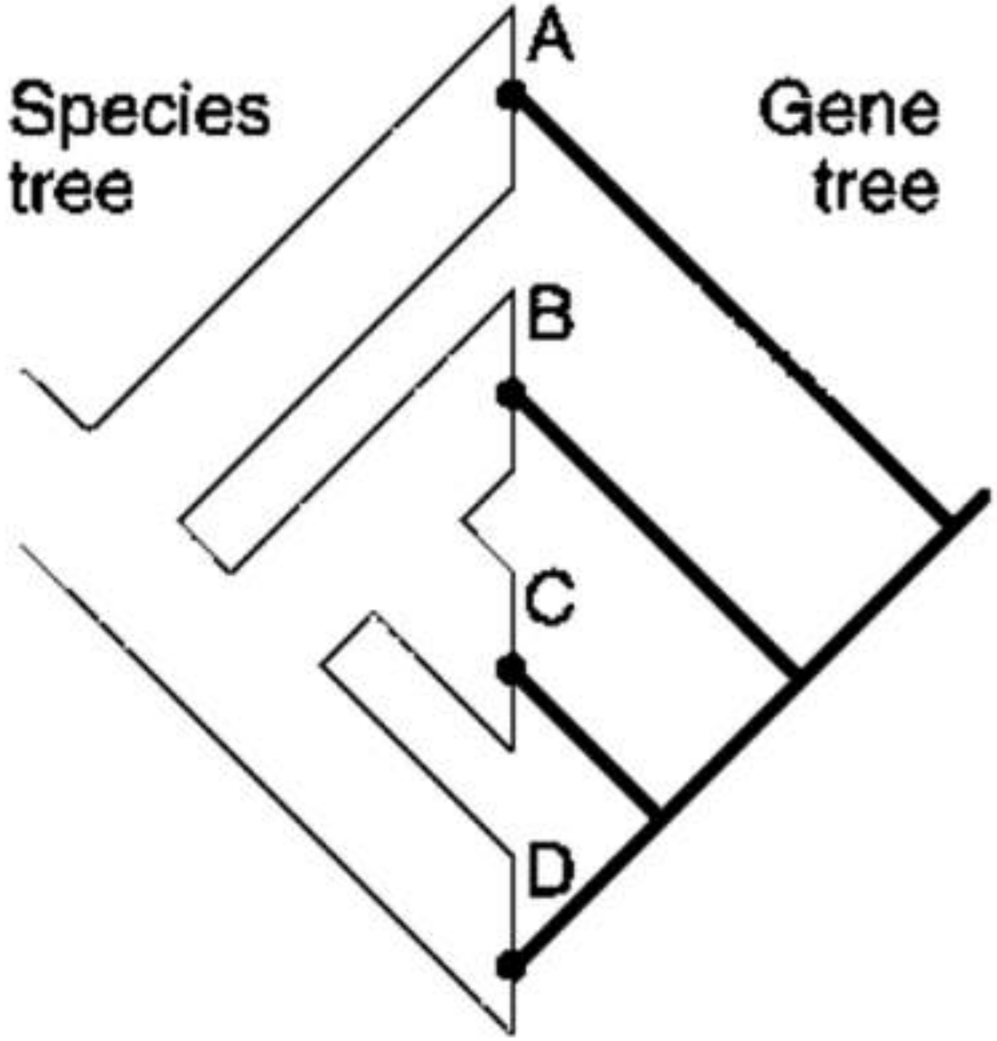


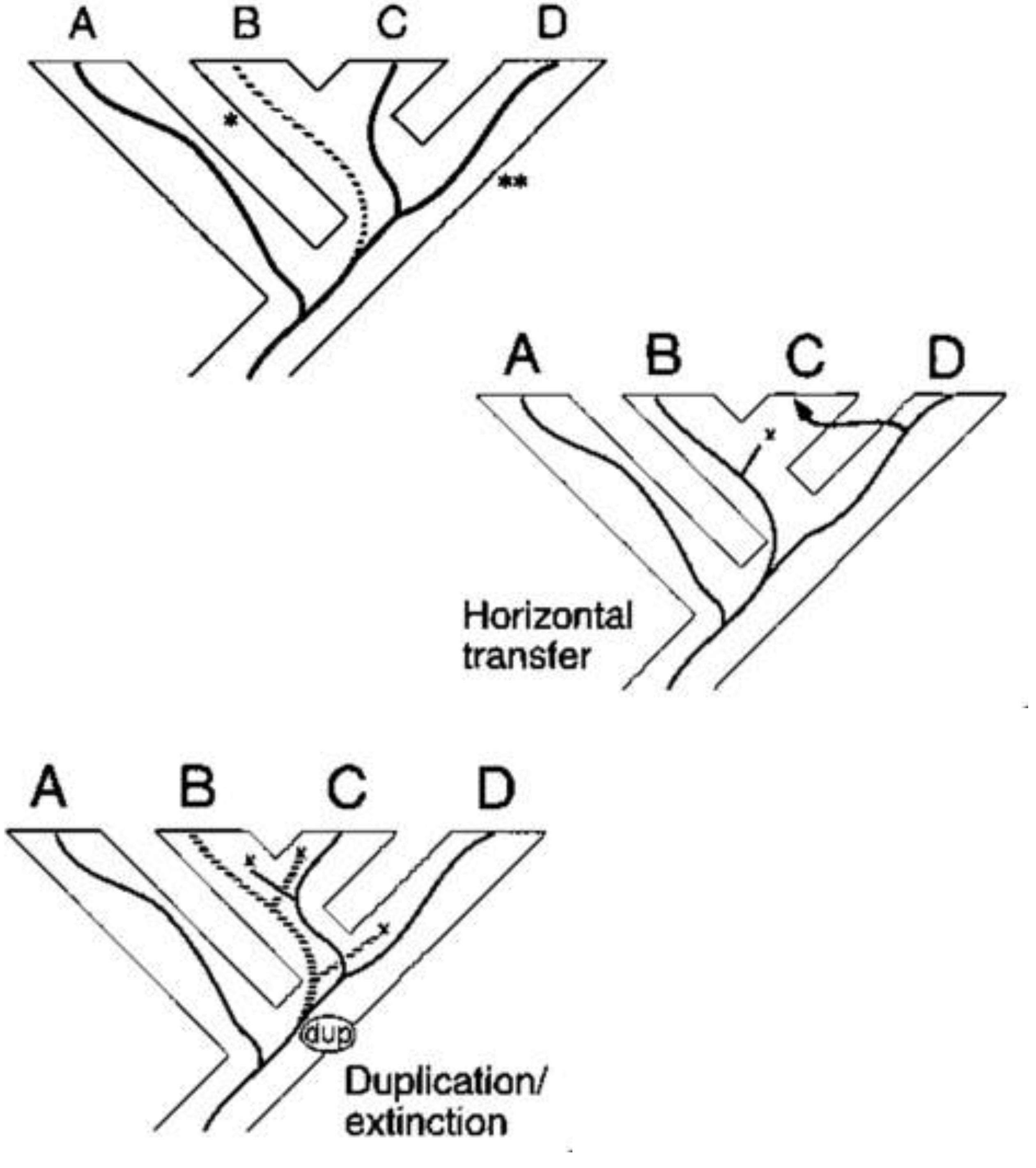
Back to the birds



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

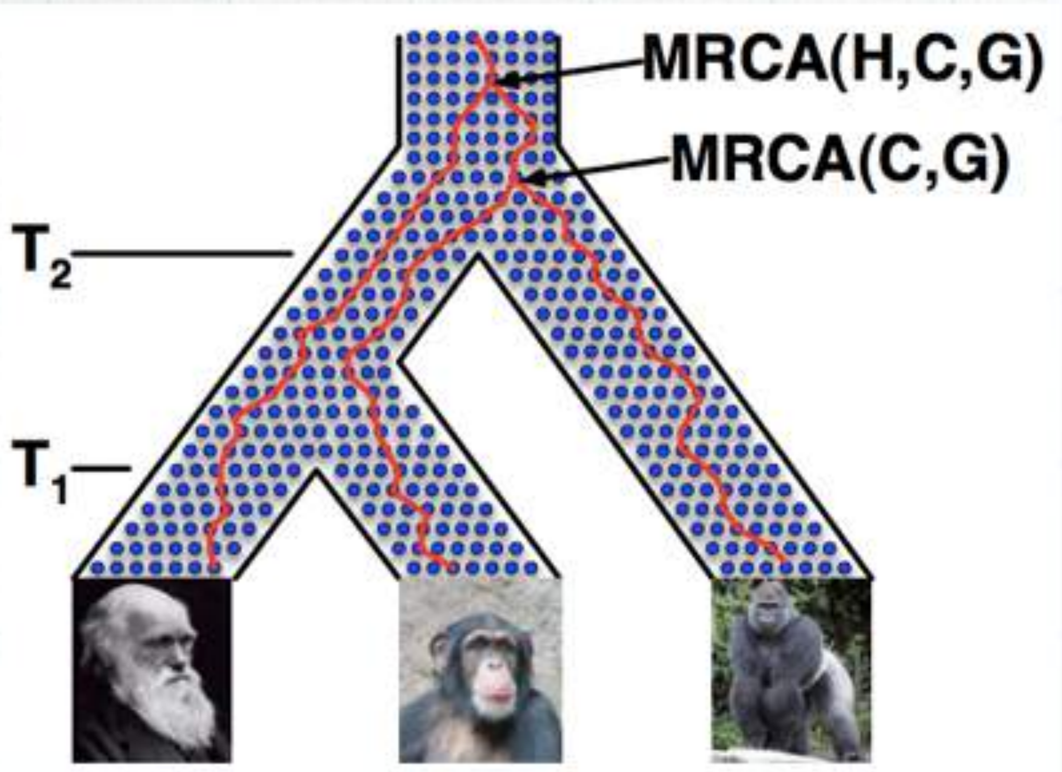
Species tree v. “gene” trees

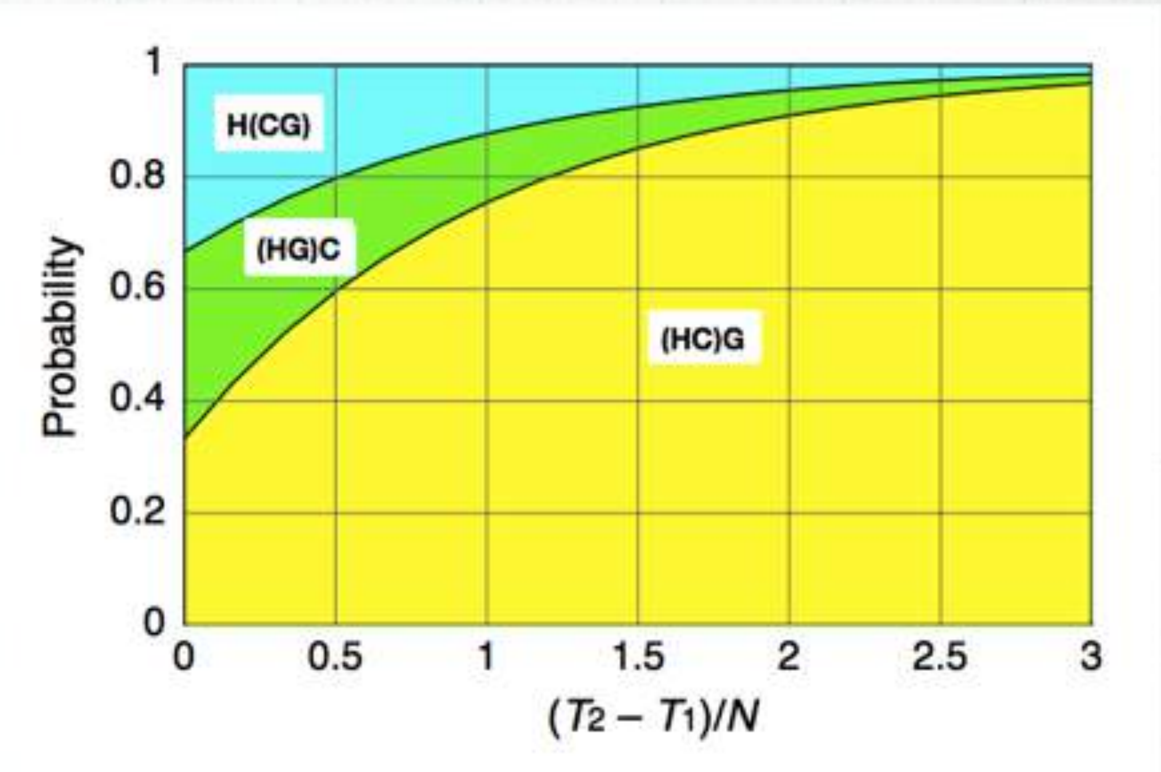




Figures from Maddison, Syst. Biol. (1997).

A source of discordance: Deep coalescence





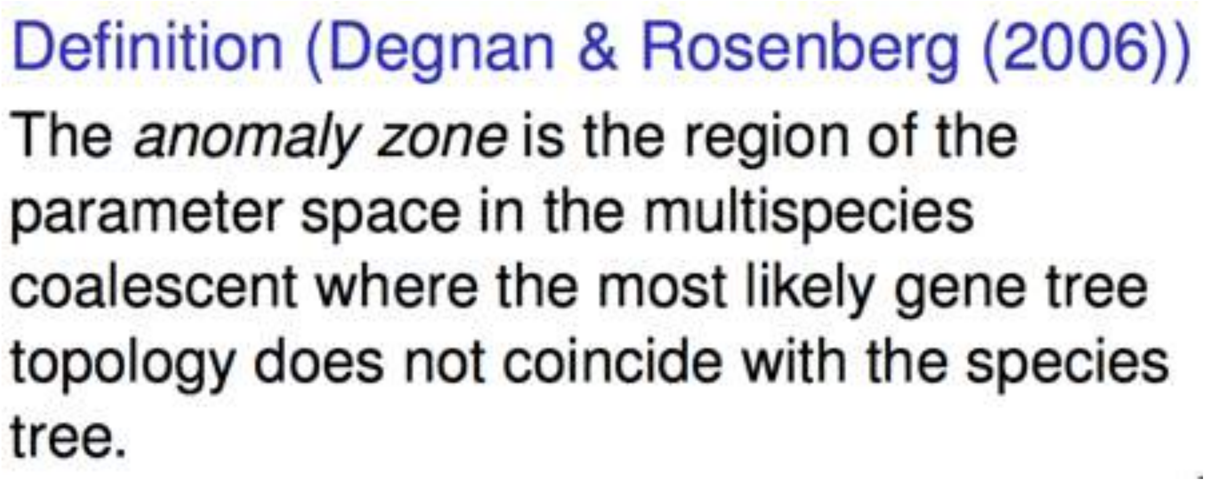
$$\mathbf{P}[((H,C),G)] = 1 - \frac{2}{3}e^{-(T_2-T_1)/N}$$

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

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

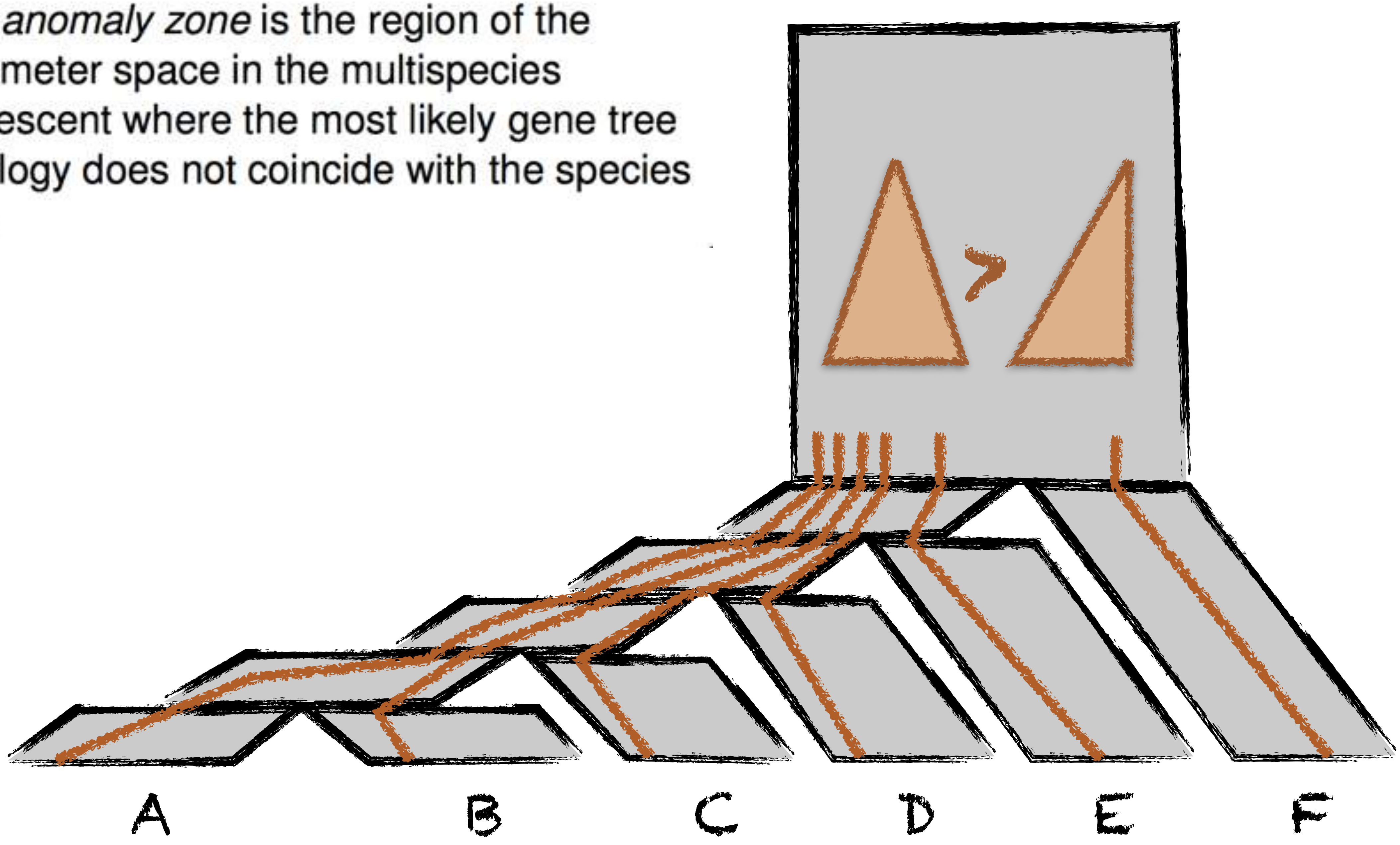
Figures by Luay Nakhleh

Anomaly zone



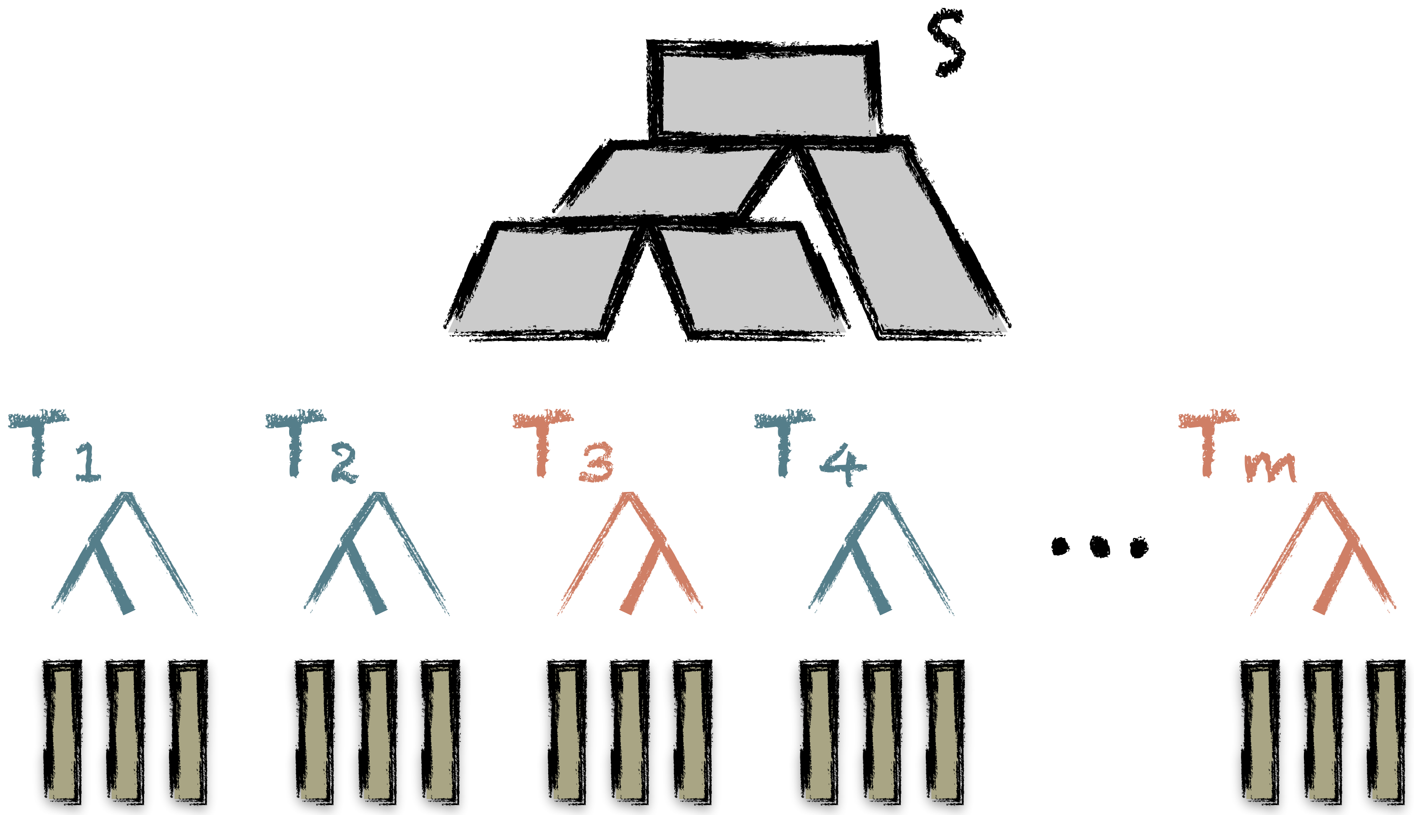
Definition (Degnan & Rosenberg (2006))

The *anomaly zone* is the region of the parameter space in the multispecies coalescent where the most likely gene tree topology does not coincide with the species tree.



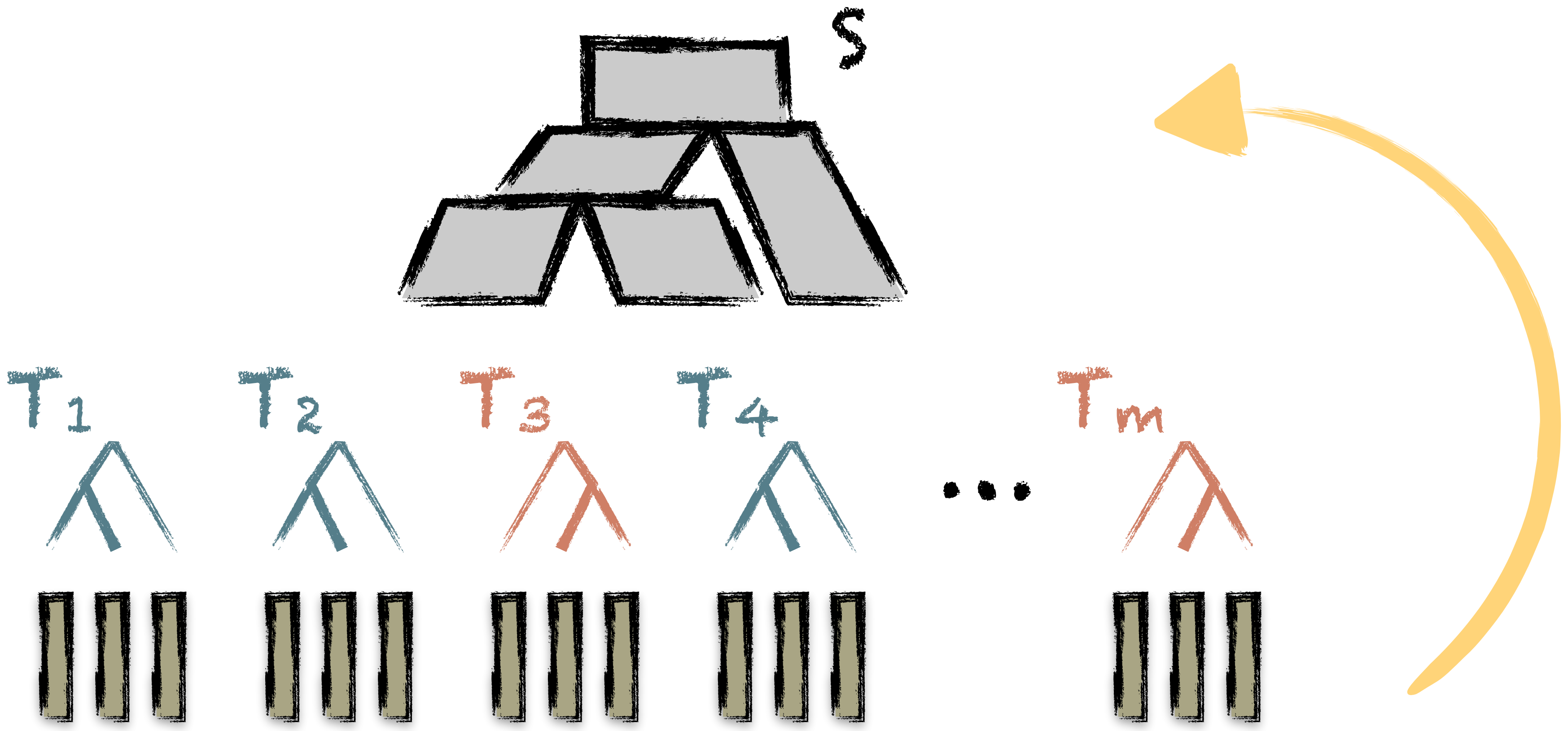
An extra layer

- Species tree: S
- Two-stage hierarchical model: 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 Markov model
- Goal: recover S from sequences

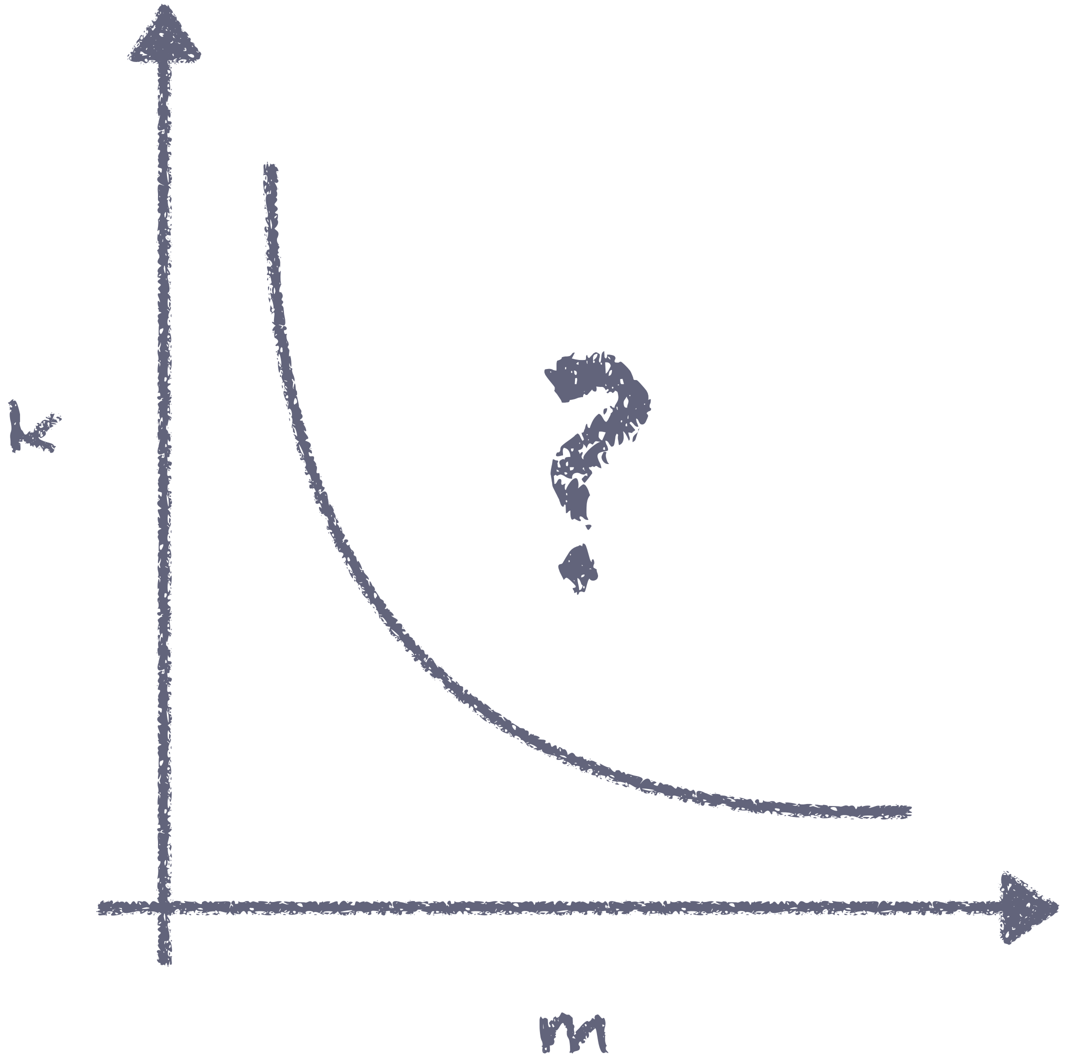


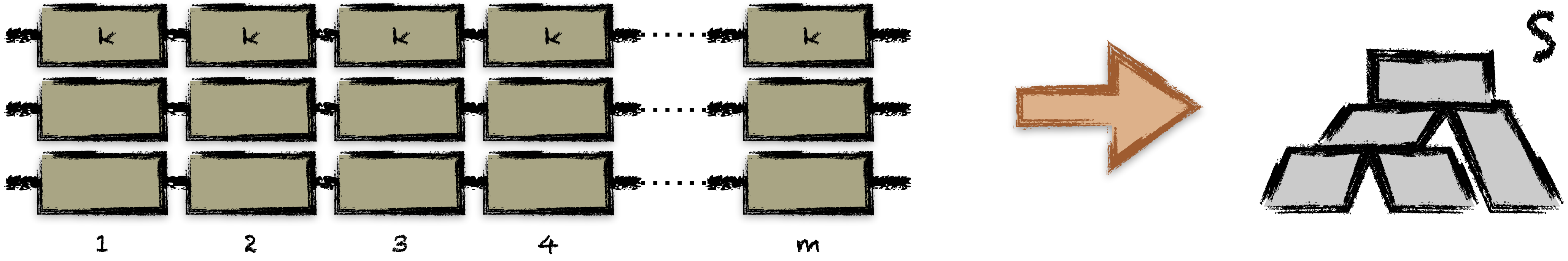
An extra layer

- Species tree: S
- Two-stage hierarchical model: 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 Markov model
- Goal: recover S from sequences



Question:
How much data is needed?

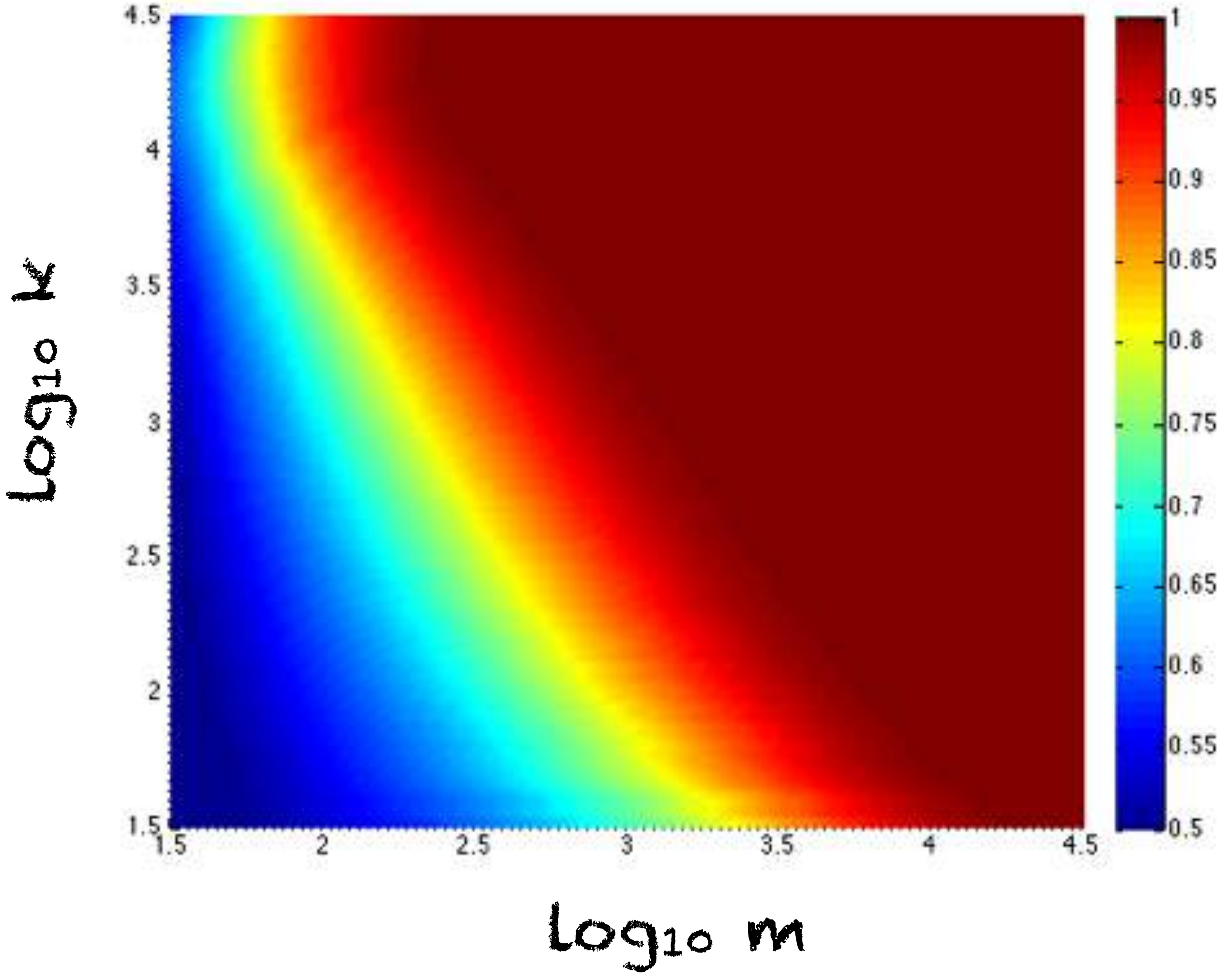




An unexpected trade-off

Theorem (Mossel & R. (2015))

Under the 3-taxon multispecies coalescent with 4-state symmetric Markov model, reconstructing the species tree requires $m = \Theta\left(f^{-2}/\sqrt{k}\right)$ *when* $k = O(f^{-2})$ *as* $f \to 0$.



A source of discordance: Horizontal gene transfer (HGT)

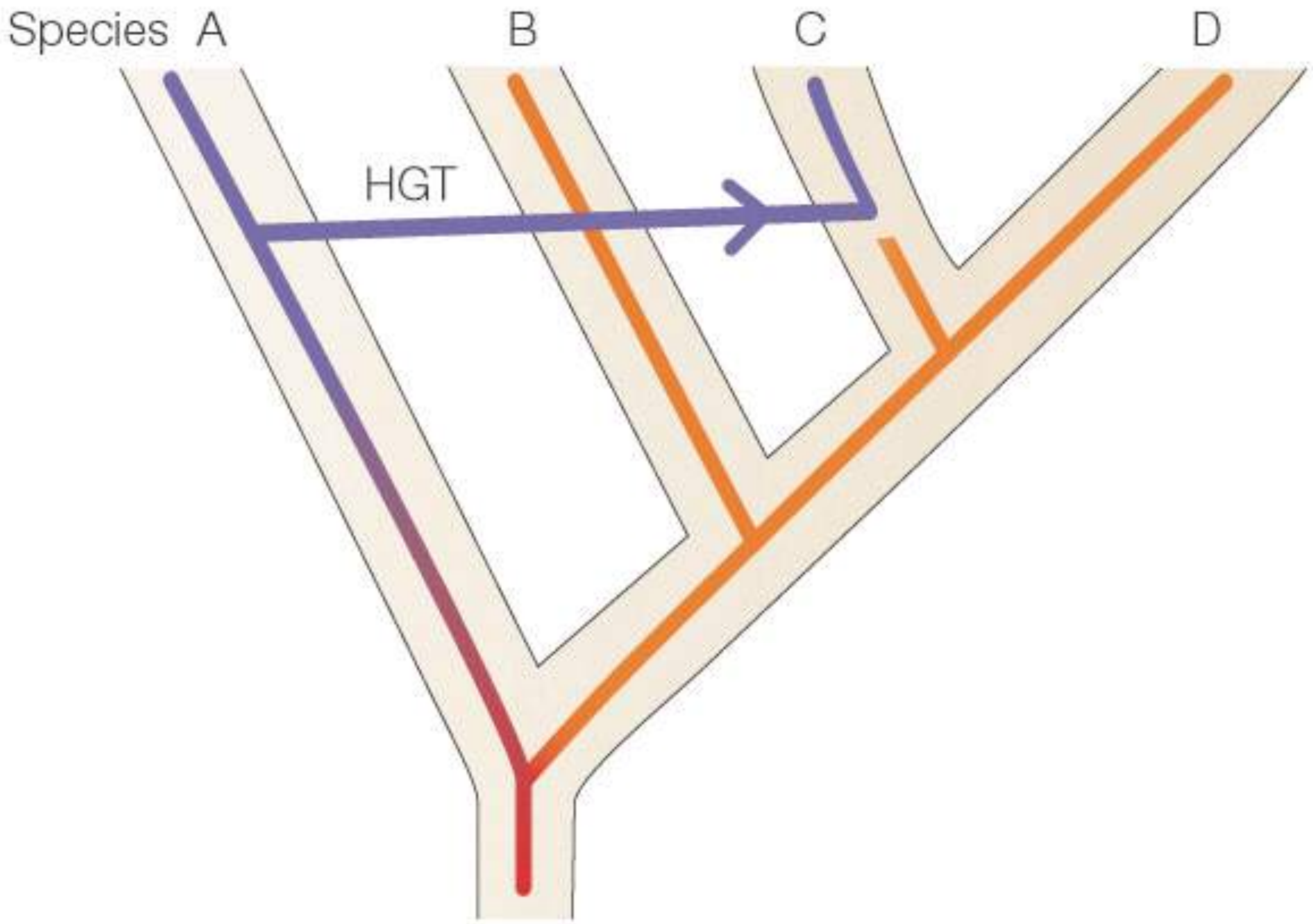
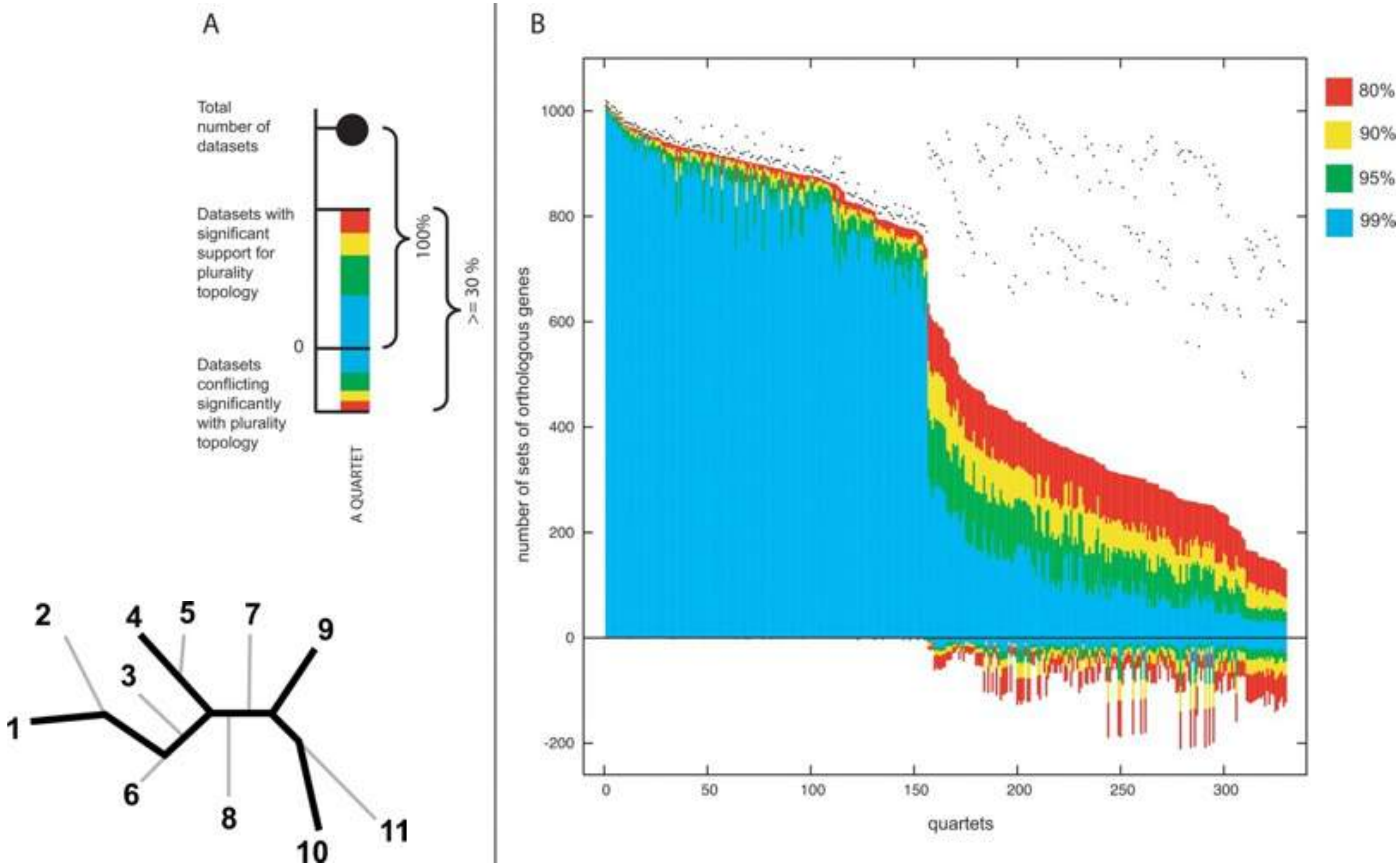


Figure from Gogarten & Townsend Nature Reviews Microbiology 3, 679-687 (2005).

Cyanobacteria



Quartet decomposition analysis of cyanobacteria.
(From: Olga Zhaxybayeva et al. Genome Res. 2006;16:1099-1108)

Subtree-prune-regraft

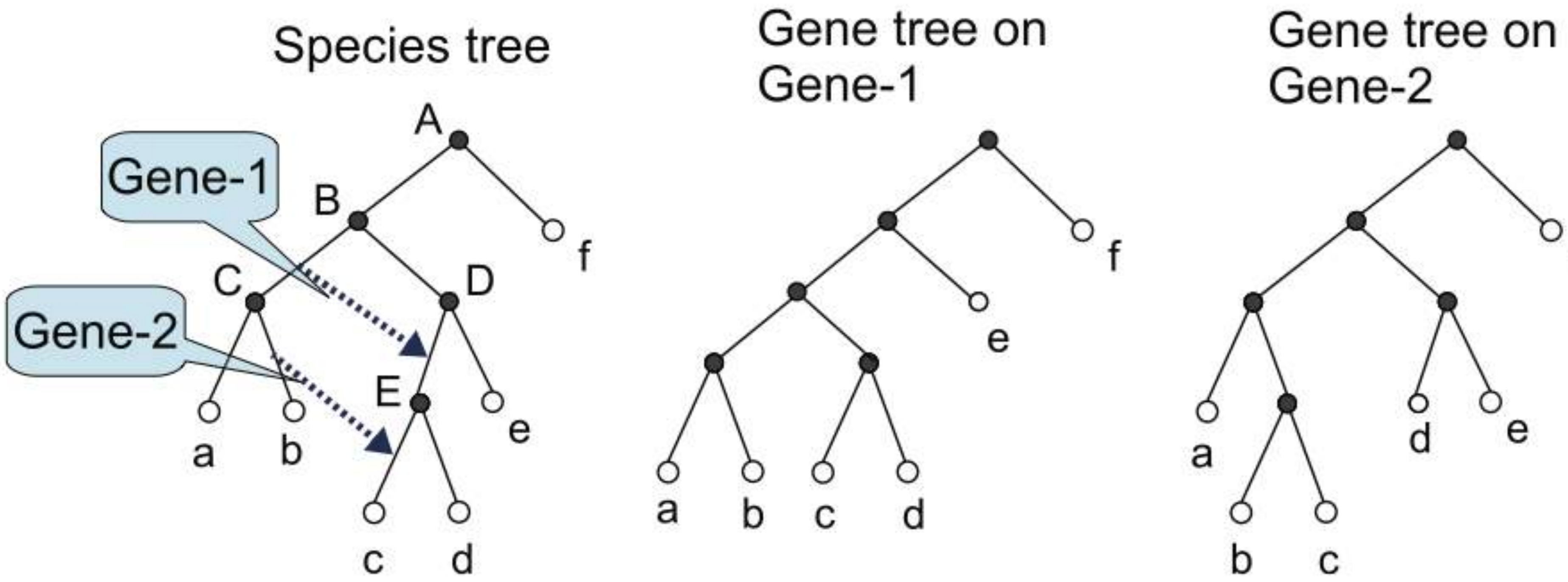
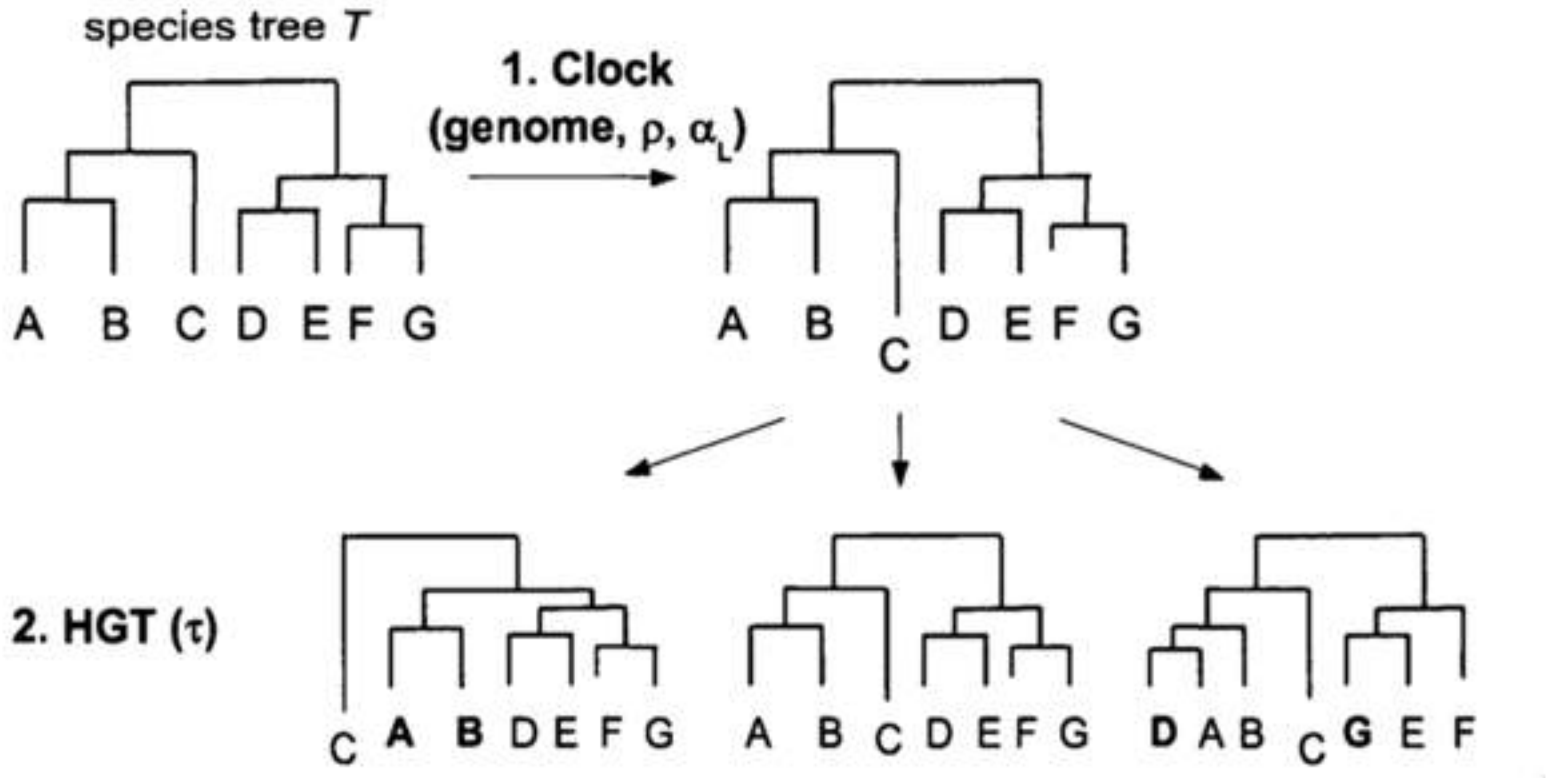


Figure from Bansal et al., Journal of Computational Biology, 8(9): 1087-1114 (2011).

HGT as combinatorial noise

- Species tree: T
- Galtier's model: for each gene g (independently and identically),
  - HGTs occur at random positions with average number ρ of HGTs per gene
  - Receivers are chosen at random among contemporaneous positions
- Goal: recover T from gene trees



Figures from Galtier, Syst. Biol. (2007).

Question:
How much HGT is too much?

Theorem (Daskalakis & R. SODA (2016))

Under Galtier's model with bounded branch lengths and a molecular clock, reconstructing the species tree from $\Omega(\log n)$ genes is possible as long as the HGT rate is constant.

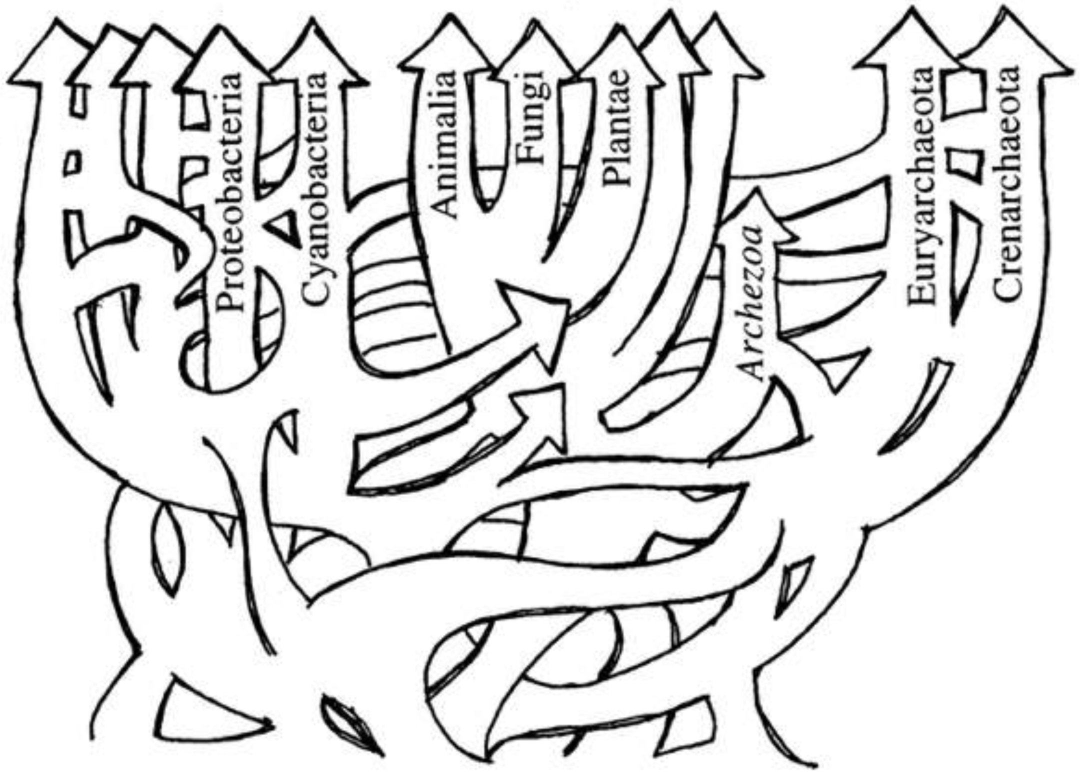
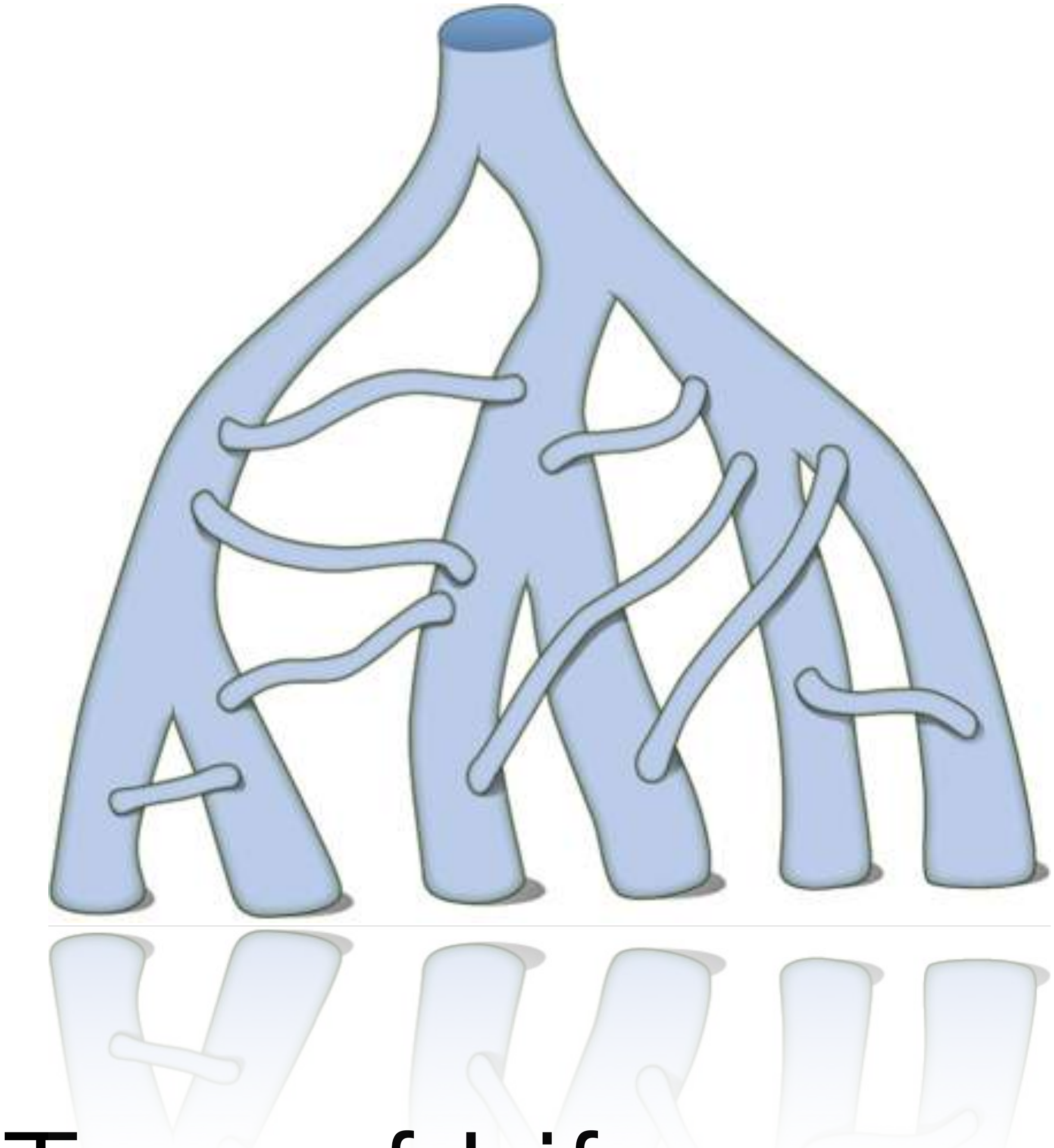


Figure from Doolittle WF (1999). Science 284:2124-2128.

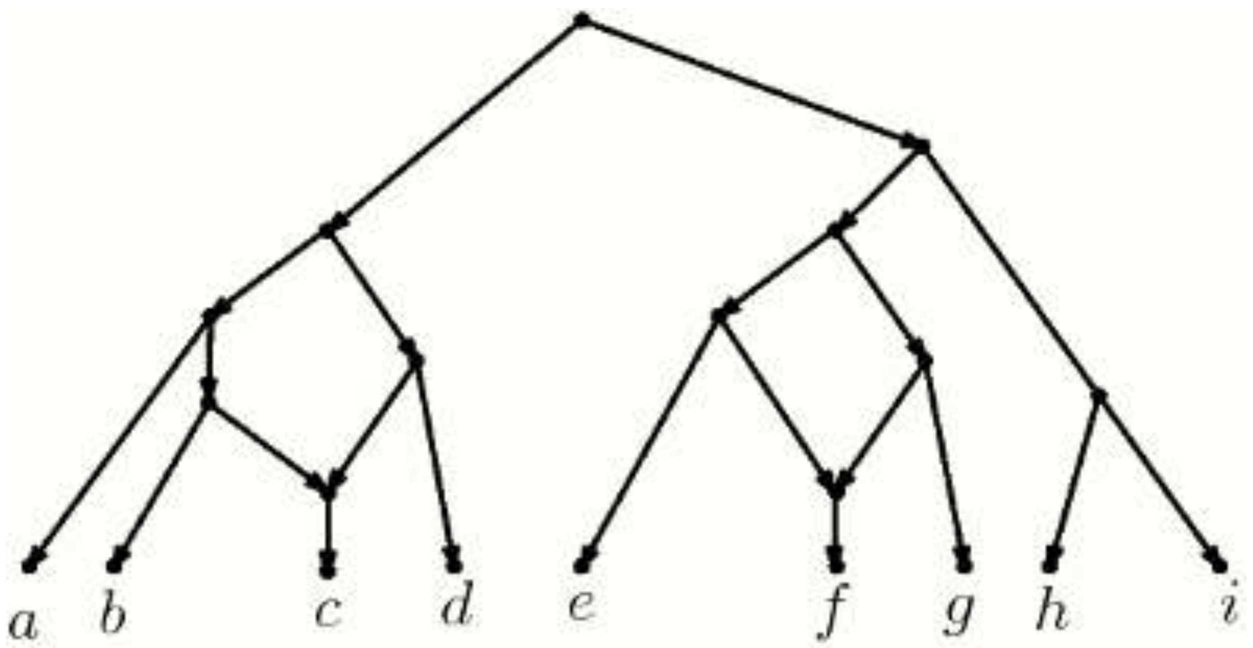


Is the Tree of Life even a tree?

Hybridization

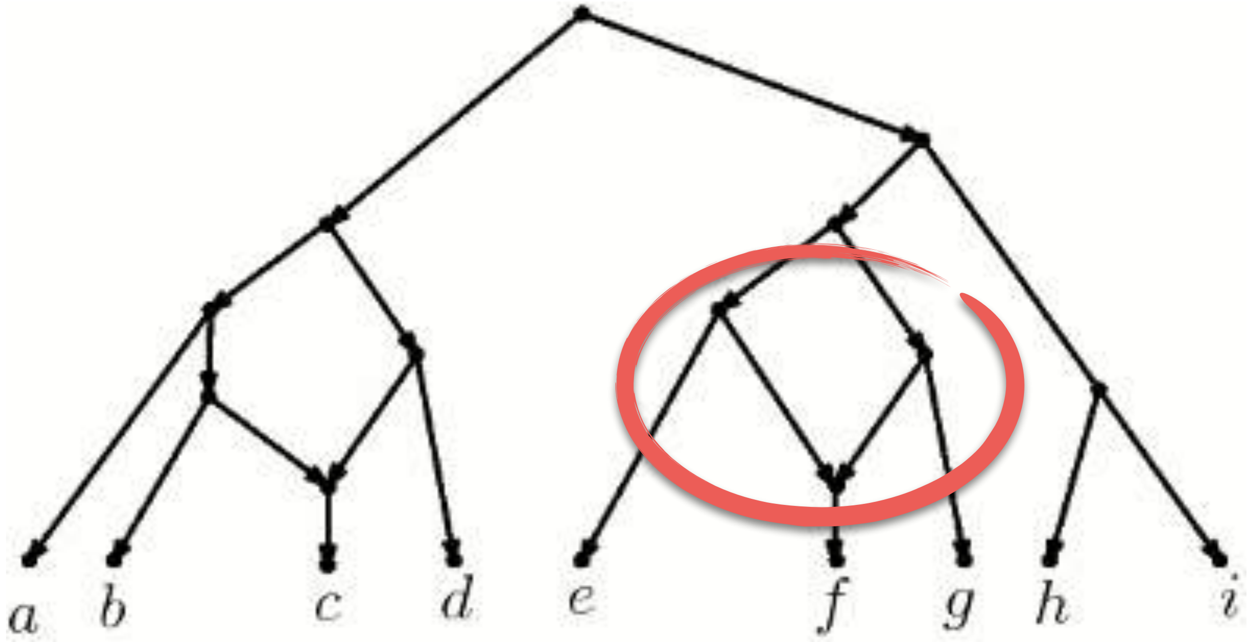
Hybridization

Beyond trees



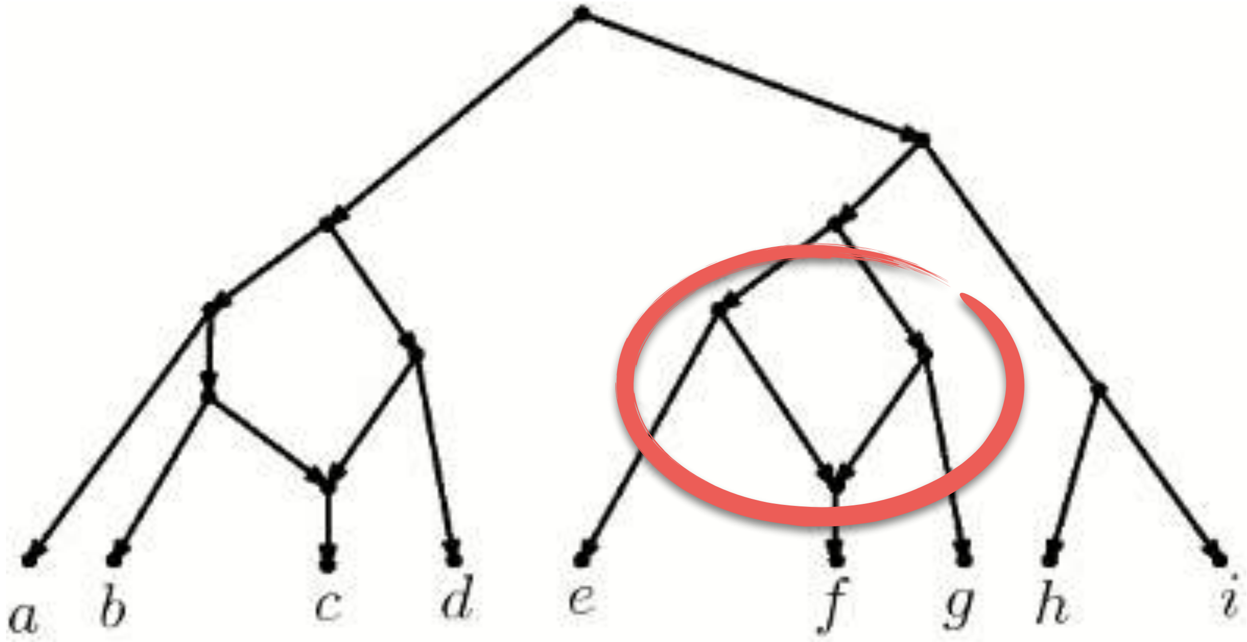
Phylogenetic network

Beyond trees



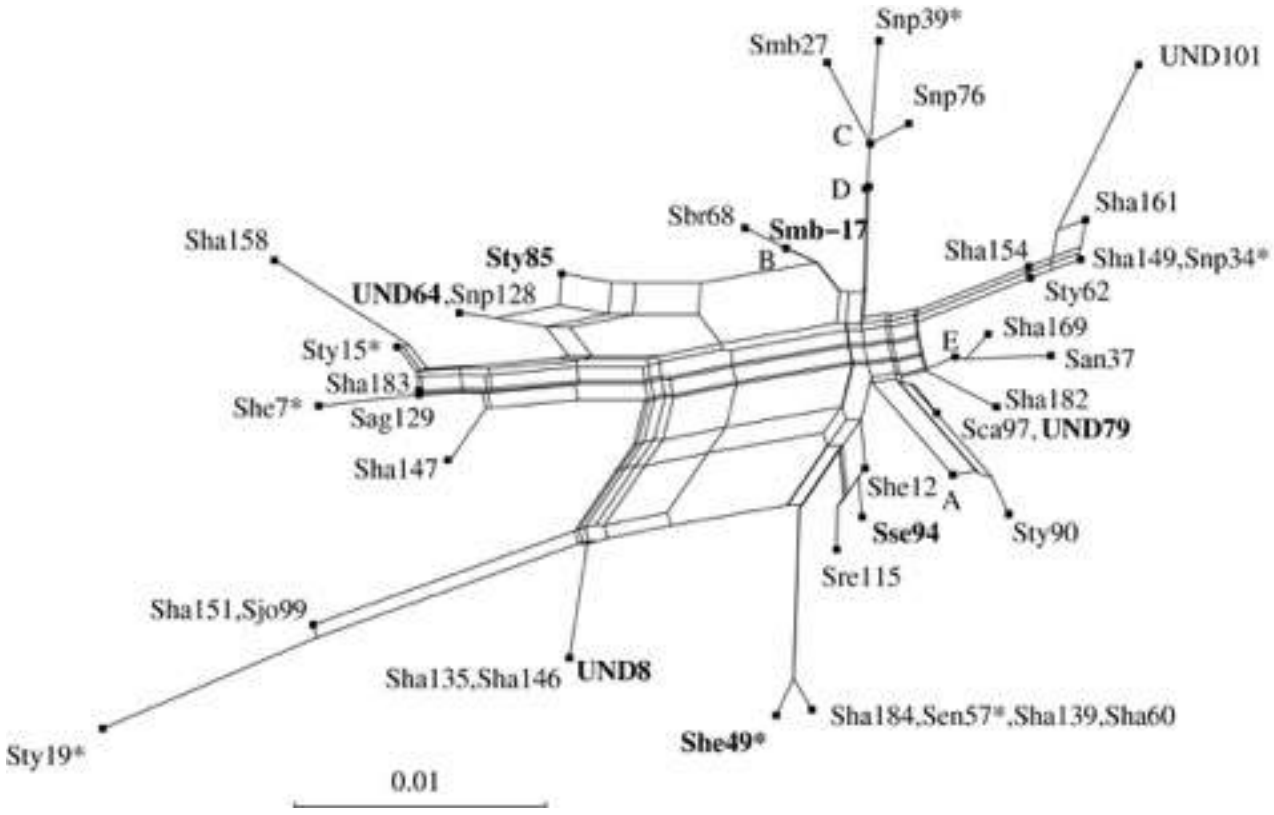
Phylogenetic network

Beyond trees



Phylogenetic network

Split network



Work supported by:

For more: http://www.math.wisc.edu/~roch/evol-gen/

Thanks

Work supported by:

For more: http://www.math.wisc.edu/~roch/evol-gen/