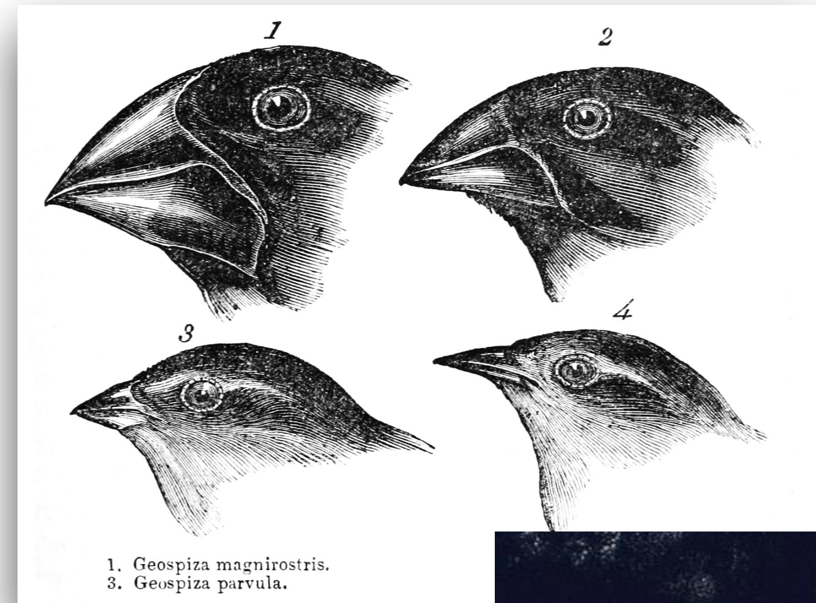
Inferring Mixtures of Trees via Multi-Site Weights

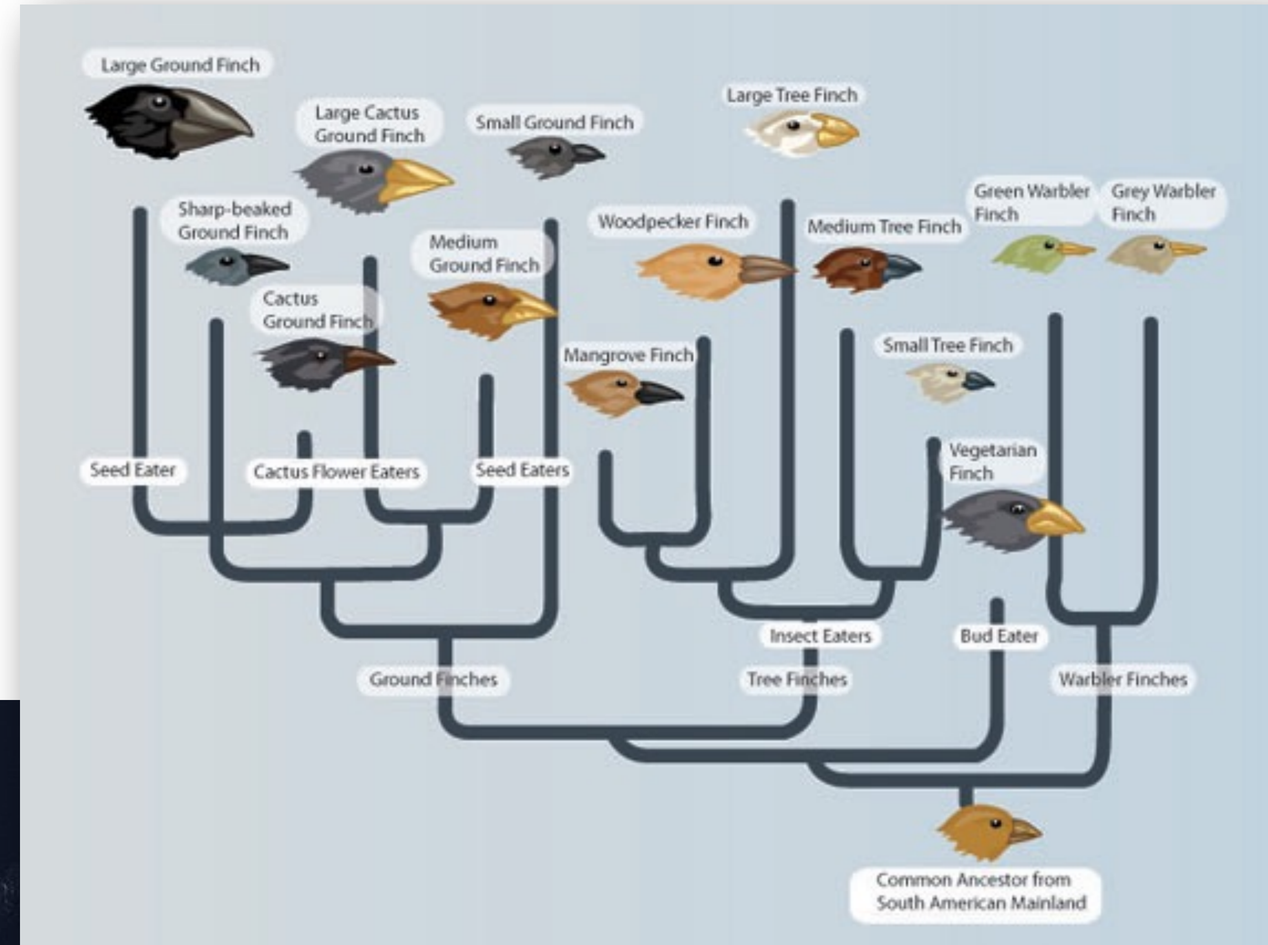
The Power of Pairs

Sébastien Roch
Department of Mathematics
University of Wisconsin-Madison

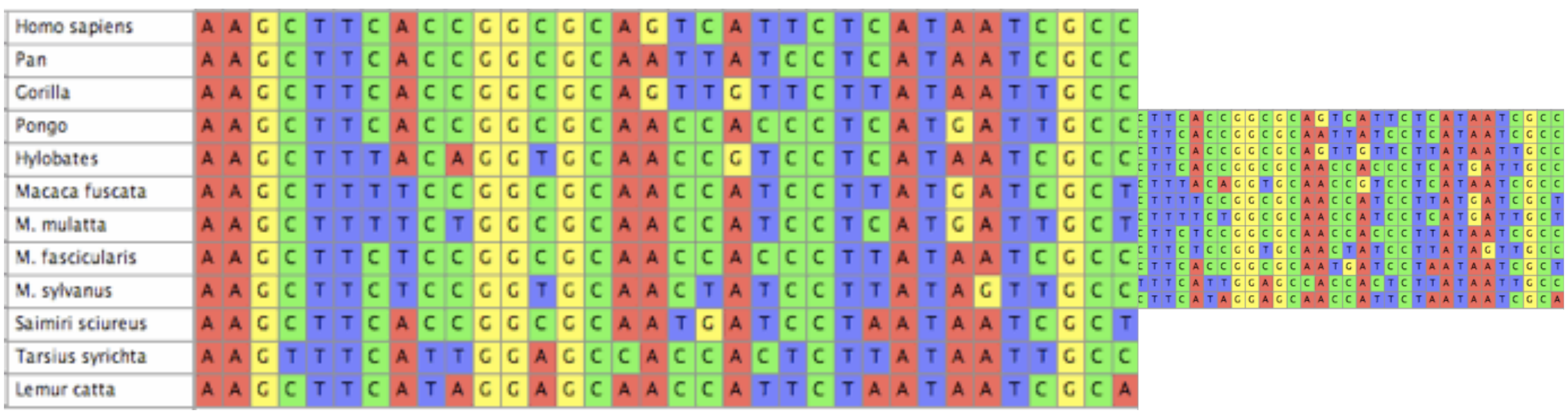
Image made with ChatGPT

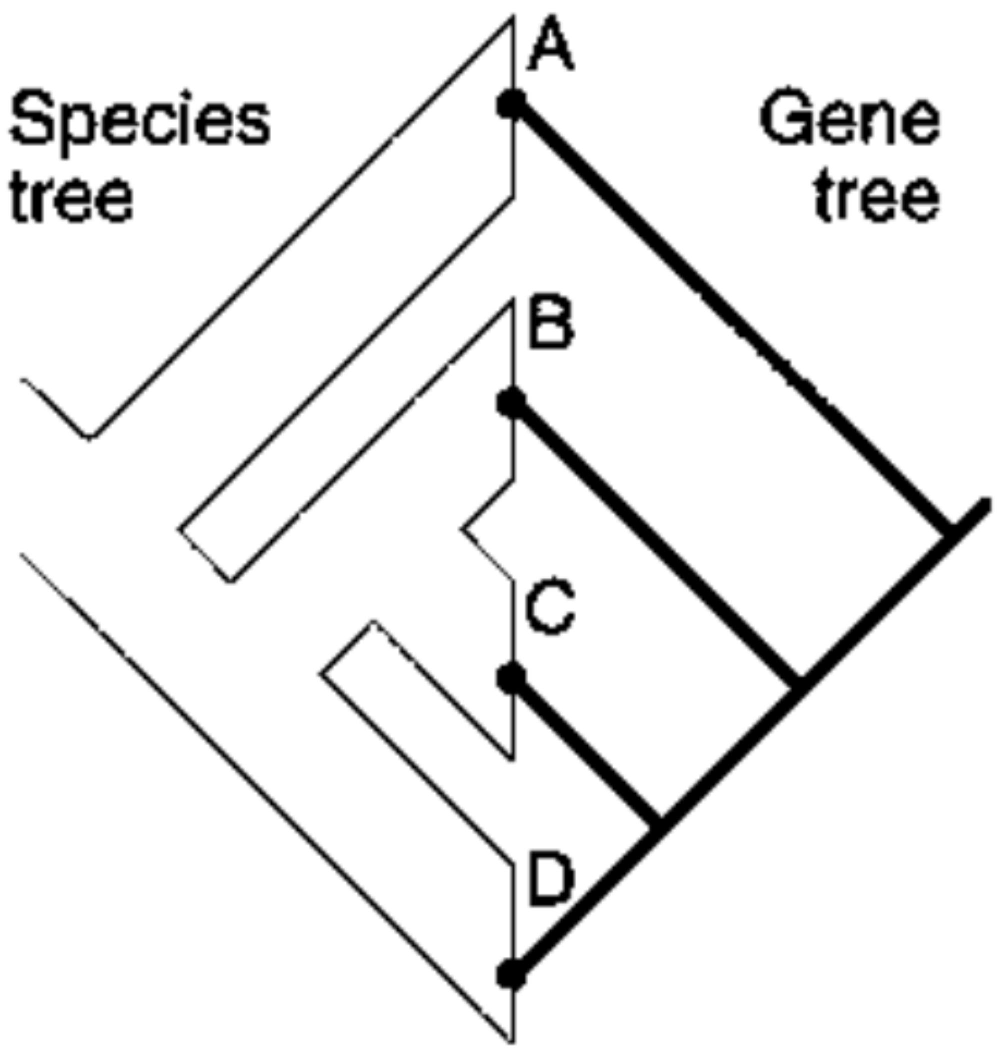
I. Background

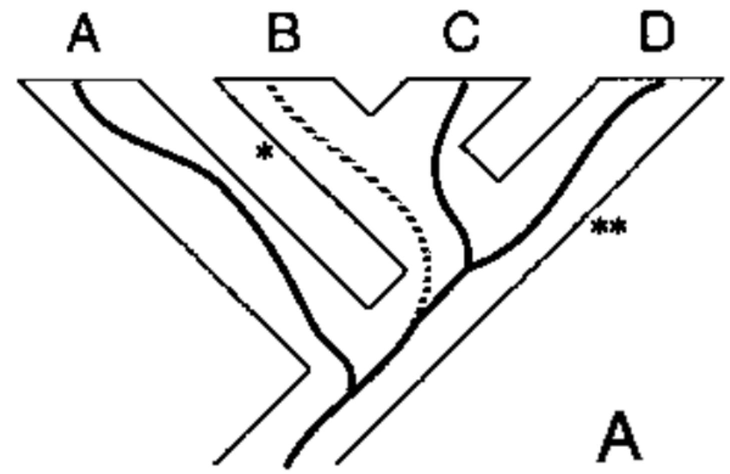


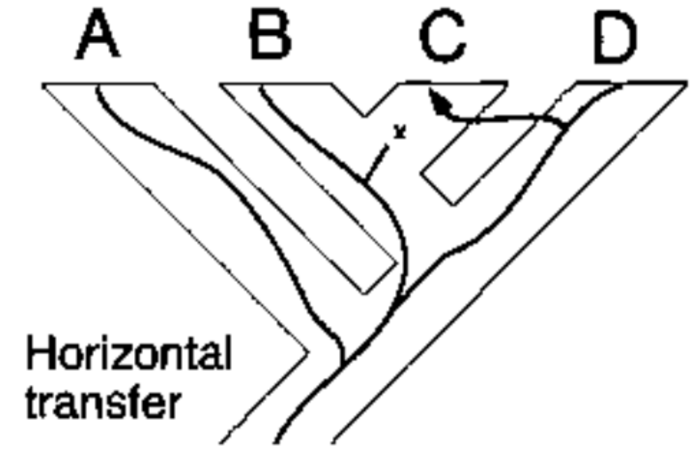


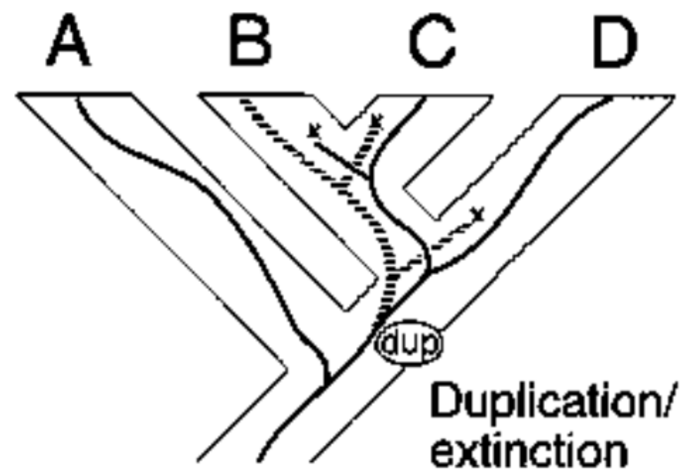
Right figure by Nature Education


Homo sapiens AAGCTTCACCGGCGCAGTCATTCTCATAATCGCC
Pan AAGCTTCACCGGCGCAATTATCCTCATAATCGCC
Gorilla AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCC
Pongo AAGCTTCACCGGCGCAACCACCCTCATGATTGCC
Hylobates AAGCTTTACAGGTGCAACCGTCCTCATAATCGCC
Macaca fuscata AAGCTTTTCCGGCGCAACCATCCTTATGATCGCT
M. mulatta AAGCTTTTCTGGCGCAACCATCCTCATGATTGCT
M. fascicularis AAGCTTCTCCGGCGCAACCACCCTTATAATCGCC
M. sylvanus AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCC
Saimiri sciureus AAGCTTCACCGGCGCAATGATCCTAATAATCGCT
Tarsius syrichta AAGTTTCATTGGAGCCACCACTCTTATAATTGCC
Lemur catta AAGCTTCATAGGAGCAACCATTCTAATAATCGCA









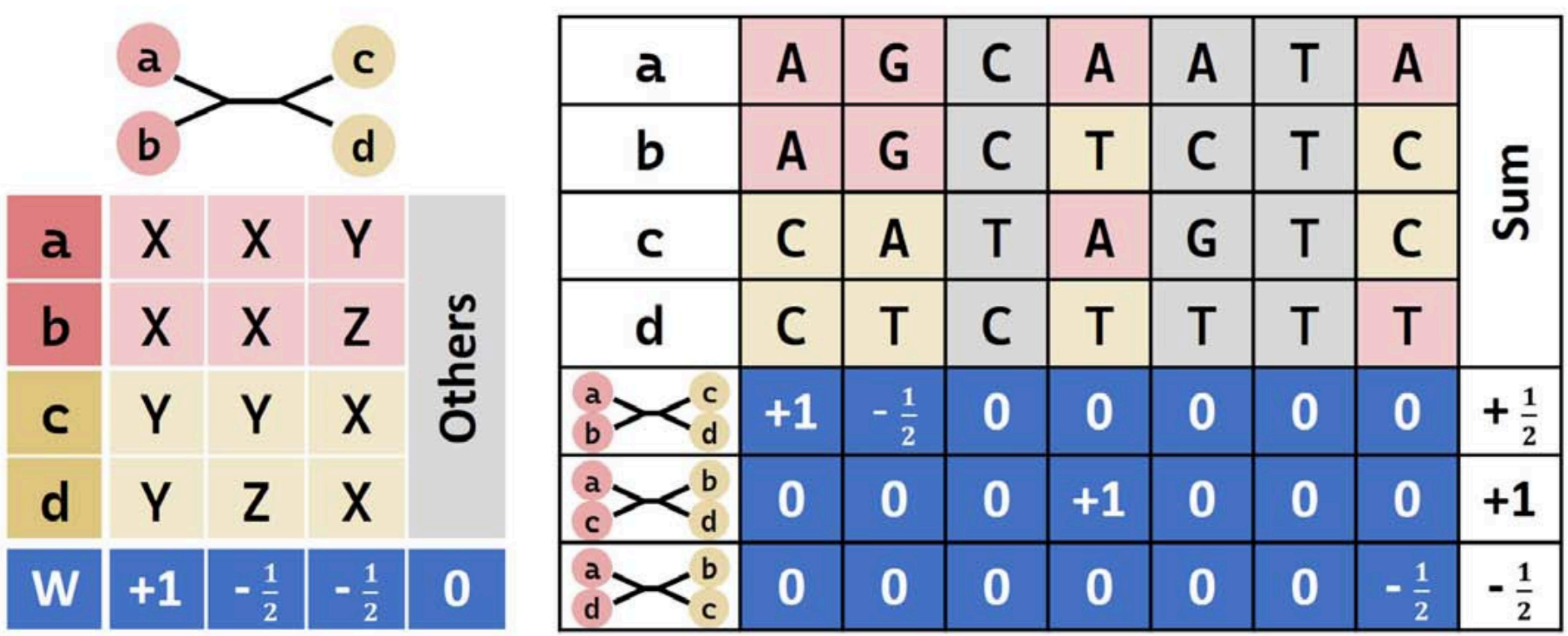
Figures from Maddison, Syst. Biol. (1997).

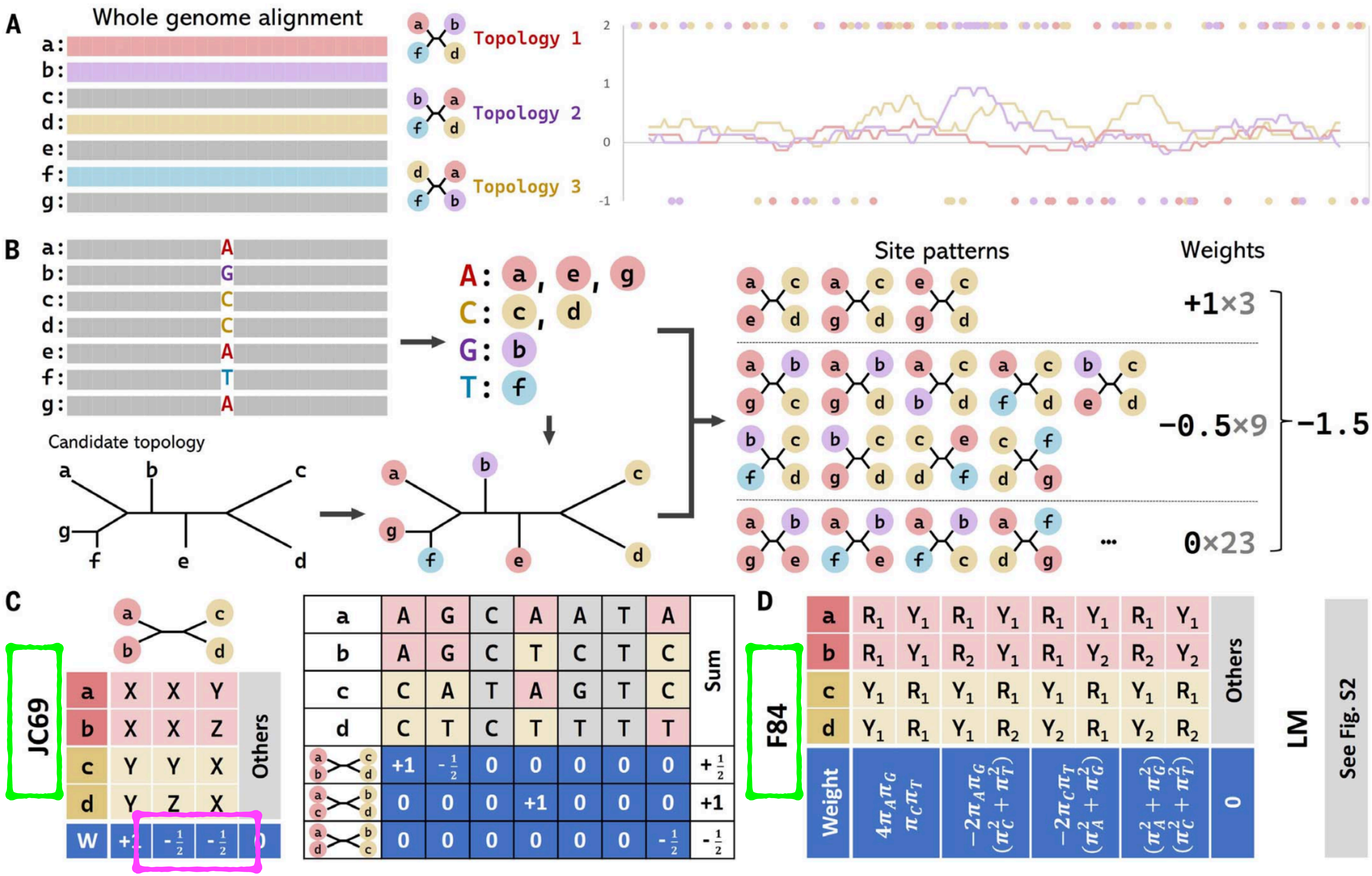
RESEARCH ARTICLE

PHYLOGENETICS

CASTER: Direct species tree inference from whole-genome alignments

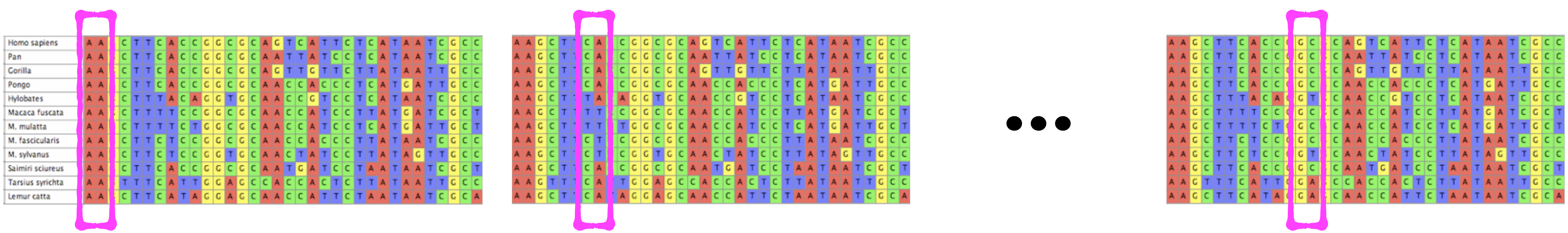
Chao Zhang[1,2,3], Rasmus Nielsen[1,3], Siavash Mirarab[4]*





A
Whole genome alignment
a: b: c: d: e: f: g:
Topology 1
Topology 2
Topology 3
2
1
0
-1
B
a: A
b: G
c: C
d: C
e: A
f: T
g: A
A: a, e, g
C: c, d
G: b
T: f
Candidate topology
Site patterns
Weights
+1×3
−0.5×9
0×23
−1.5
C
JC69
W
Others
Sum
D
F84
Weight
Others
LM
See Fig. S2

a	RN	RN	YN	YN	NR	NR	NY	NY	RN	YN	NN	NN
b	YN	YN	RN	RN	NY	NY	NR	NR	YN	RN	NN	NN
c	NR	NY	NR	NY	RN	YN	RN	YN	NN	NN	RN	YN
d	NY	NR	NY	NR	YN	RN	YN	RN	NN	NN	YN	RN
w	$+1$								$-4\pi_R\pi_Y$			

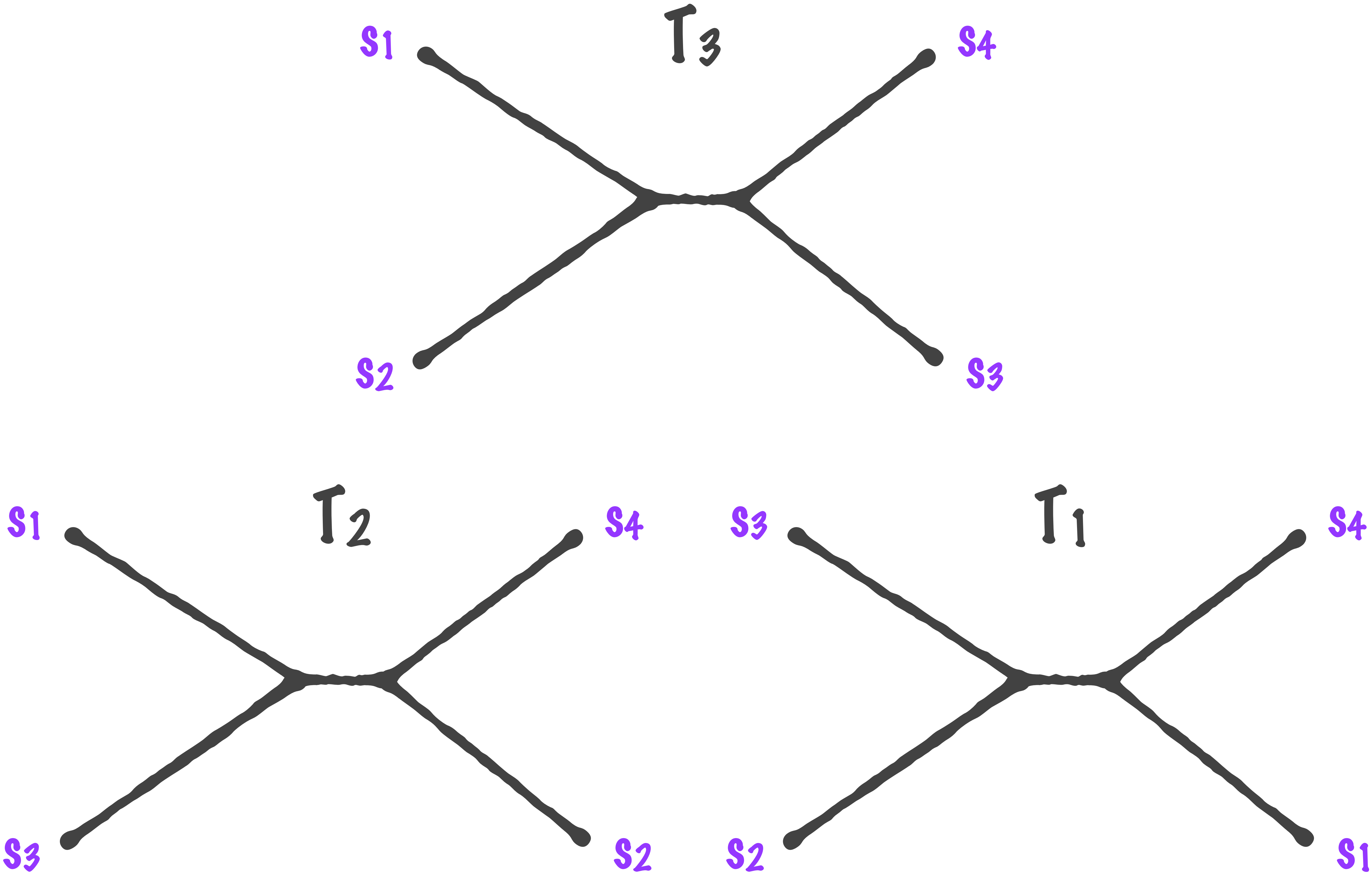


Proposition 2. *For the true gene tree* $\mathbf{G}_i$ *of four leaves with topology* $ab|cd$ *(irrespective of the species tree topology), let* l_a, l_b, l_c, l_d *denote the terminal branch lengths, and* l_x *denote the internal branch length in substitution units, then*

$$\mathbb{E}\left[w_i(ab|cd)\middle|\mathbf{G}_i\right] - \gamma e^{-\alpha(l_a+l_b+l_c+l_d)}\left(1 - e^{-\beta l_x}\right) = \mathbb{E}\left[w_i(ac|bd)\middle|\mathbf{G}_i\right] \tag{S21}$$

for some $\alpha > 0$, $\beta > 0$, *and* $\gamma > 0$.

II. Linear Tests for Mixtures



T3
S1
S4
S2
S3
T2
S1
S4
S3
S2
T1
S3
S4
S2
S1

Evolutionary Process

- States evolve in space $[\ell] = \{1, \ldots, \ell\}$
- Governed by reversible rate matrix Q with stationary distribution π (i.e., $\pi_i Q_{ij} = \pi_j Q_{ji}$)
- For topology T_i and branch lengths $\vec{t}$: distribution $P_{T_i}(\vec{t})$

For topology T_3 with internal nodes r_1, r_2 and branch lengths $(t_0, t_1, t_2, t_3, t_4)$:

$$P_{T_3}(\vec{t})(w, x, y, z) = \sum_{u,v \in [\ell]} \pi_u (e^{Qt_0})_{uv} (e^{Qt_1})_{uw} (e^{Qt_2})_{ux} (e^{Qt_3})_{vy} (e^{Qt_4})_{vz}$$

where (w, x, y, z) are observed states at leaves (s_1, s_2, s_3, s_4) and (u, v) are unobserved states at the internal nodes (r_1, r_2).

Definition: Single-Tree Mixture Distribution

A **single-tree mixture distribution** on a topology T is a probability distribution μ_T on $[\ell]^4$ that is a convex combination of distributions generated on T with different branch lengths:

$$\mu_T = \sum_{k=1}^{N} c_k P_T(\vec{t}_k)$$

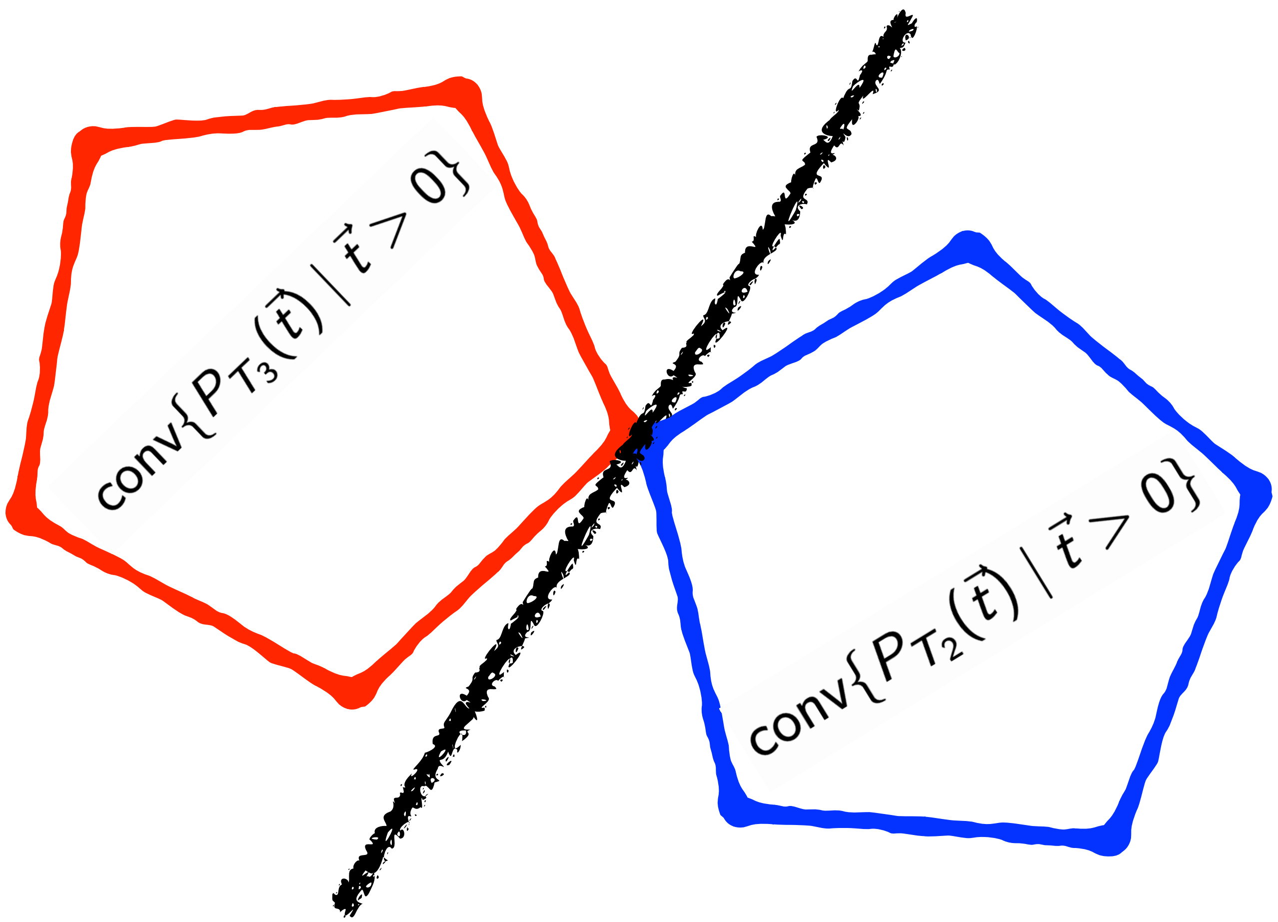
where:

- $c_k > 0$ and $\sum_{k=1}^{N} c_k = 1$ (convex combination)
- Each $\vec{t}_k$ is a vector of positive branch lengths

Definition [Stefankovic-Vigoda'07]

A **linear test** for distinguishing topology T_3 from T_2 is a real-valued function $H : [\ell]^4 \to \mathbb{R}$ such that:

$$\mathbb{E}_{\mu_{T_3}}[H] > 0 \quad \text{for any mixture } \mu_{T_3} \text{ on topology } T_3$$
$$\mathbb{E}_{\mu_{T_2}}[H] < 0 \quad \text{for any mixture } \mu_{T_2} \text{ on topology } T_2$$



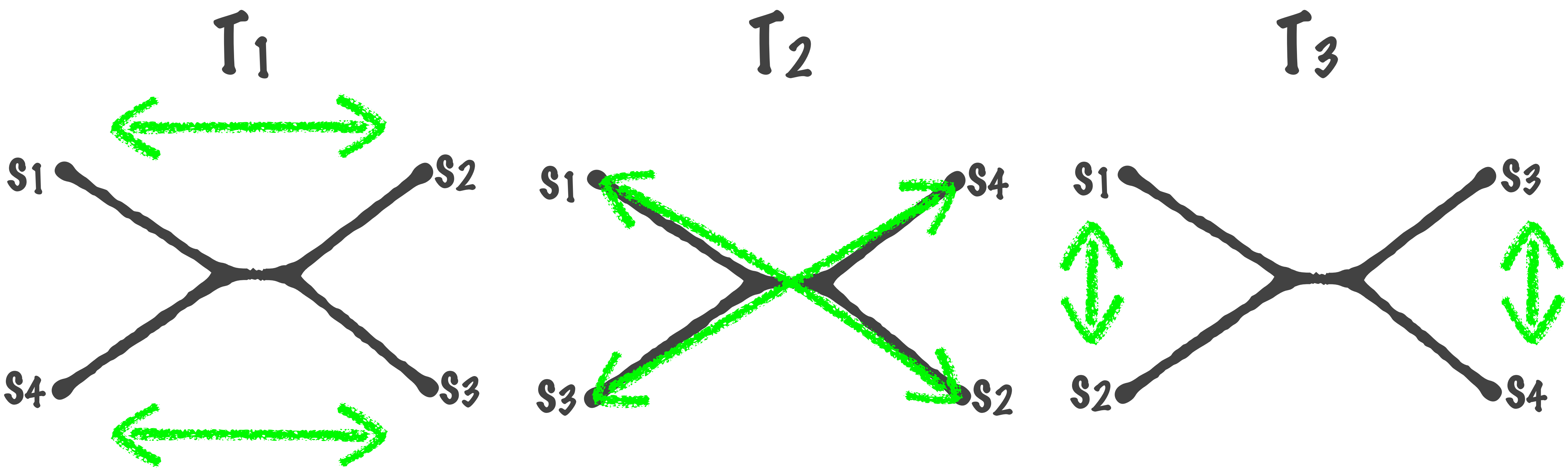
Hyperdimensional Oranges (Kim'00)

Consider the group $R = \{e, (1\,2)(3\,4), (1\,3)(2\,4), (1\,4)(2\,3)\}$. For any $g \in R$, we have $T_i^g = T_i$ for $i = 1, 2, 3$.

Proposition [Stefankovic-Vigoda'07]

If a linear test H for distinguishing topology T_3 from T_2 exists, then an R-invariant linear test for T_3 vs T_2 also exists.

e.g., $H(w,x,y,z) = H(x,w,z,y)$



For JC69, it is shown in [Stefankovic-Vigoda'07] that:

- There is a unique R-invariant linear test (up to scaling) for distinguishing T_3 from T_2 that is also invariant under any permutation of the states
- That test is a linear invariant [Lake'87]: $\mathbb{E}_{\mu_{T_1}}[H] \equiv 0$
- It coincides with the CASTER weights
- For the more general TN93 model, linear (topology) invariants were derived in [Casanellas-Homs-Torres'24]

For K3P, it is shown in [Stefankovic-Vigoda'07] that:

- There are no such tests (see also [Sturmfels-Sullivant'05])

III. A Mathematical Framework

Definition [R.'25]

A **linear score** for distinguishing topology T_3 from T_2 and T_1 is a real-valued function $H : [\ell]^4 \to \mathbb{R}$ such that: for any mixtures μ_{T_1}, μ_{T_2}, μ_{T_3} on T_1, T_2, T_3 respectively

$$\mathbb{E}_{\mu_{T_3}}[H] > \mathbb{E}_{\mu_{T_2}}[H], \qquad \mathbb{E}_{\mu_{T_3}}[H] > \mathbb{E}_{\mu_{T_1}}[H].$$

Furthermore, we require

$$\mathbb{E}_{\mu_{T_3}}[H] \geq 0, \quad \mathbb{E}_{\mu_{T_2}}[H] \leq 0, \quad \mathbb{E}_{\mu_{T_1}}[H] \leq 0.$$

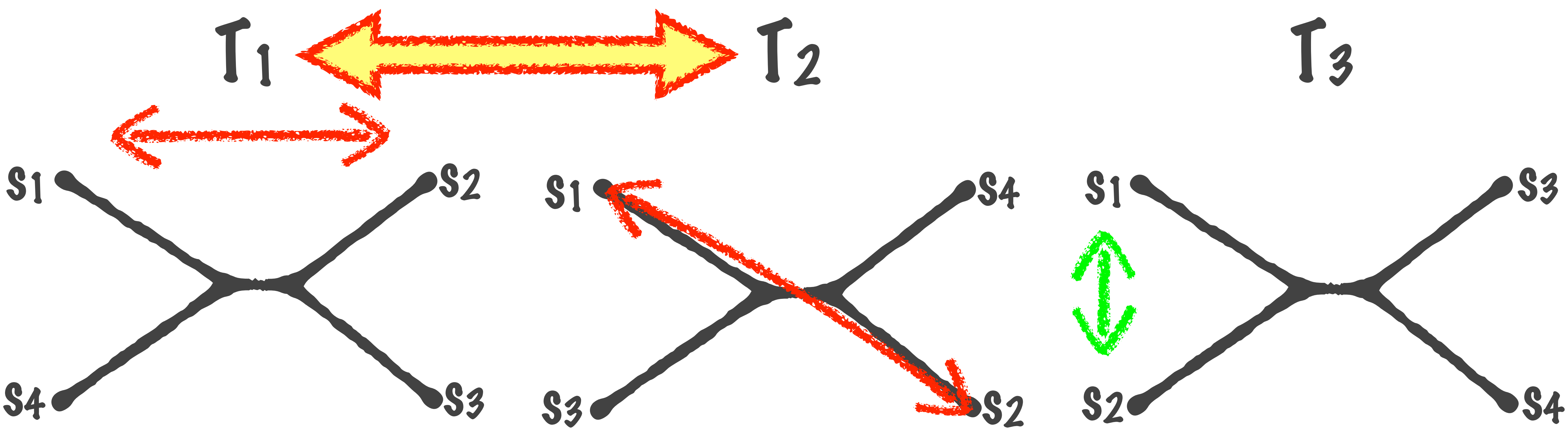
Consider the group

$$K = \{e, (1\,2), (3\,4), (1\,2)(3\,4), (1\,3)(2\,4), (1\,4)(2\,3), (1\,3\,2\,4), (1\,4\,2\,3)\}.$$

For any $g \in K$, we have $T_3^g = T_3$. Does not hold for T_2 and T_1: e.g., $(1\,2)$ swaps T_2 and T_1.

Proposition [R.'25]

If a linear score H for distinguishing topology T_3 from T_2 and T_1 exists, then a K-invariant linear score also exists.



π-Weighted Inner Product

Let π be the stationary distribution of the reversible rate matrix Q. For functions $f, g : [\ell] \to \mathbb{R}$, define the weighted inner product:

$$\langle f, g \rangle_\pi := \sum_{i=1}^{\ell} \pi_i f(i) g(i)$$

Because Q is self-adjoint in the π-weighted inner product, there exists an orthonormal eigenbasis $\{\varphi_1, \dots, \varphi_\ell\}$ with:

- $Q\varphi_a = \lambda_a \varphi_a$ and $\langle \varphi_a, \varphi_b \rangle_\pi = \delta_{ab}$
- $\lambda_1 = 0$ with $\varphi_1(i) = 1$ for all $i \in [\ell]$ (constant function)

A related inner product was used in [Casanellas-Homs-Torres'24].

- Identify four-variable functions $h : [\ell]^4 \to \mathbb{R}$ with vector space

$$U := \mathbb{R}^{[\ell]} \otimes \mathbb{R}^{[\ell]} \otimes \mathbb{R}^{[\ell]} \otimes \mathbb{R}^{[\ell]} \cong \mathbb{R}^{[\ell]^4}$$

- Simple tensors

$$f_1 \otimes f_2 \otimes f_3 \otimes f_4 : (w, x, y, z) \mapsto f_1(w) f_2(x) f_3(y) f_4(z)$$

- A basis of U

eigenfunction of Q

$$\{\varphi_a \otimes \varphi_b \otimes \boxed{\varphi_c} \otimes \varphi_d : (a, b, c, d) \in [\ell]^4\}$$

that is orthonormal under the inner product

$$\langle h, k \rangle_{\pi^{\otimes 4}} = \sum_{w,x,y,z \in [\ell]} \pi_w \pi_x \pi_y \pi_z \, h(w, x, y, z) \, k(w, x, y, z)$$

- K-invariant subspace

$$U^K = \{H \in U \mid g \cdot H = H \text{ for all } g \in K\}$$

where $g \cdot H(a_1, a_2, a_3, a_4) = H(g \cdot (a_1, a_2, a_3, a_4))$ and $g \cdot (a_1, a_2, a_3, a_4) = (a_{g^{-1}(1)}, a_{g^{-1}(2)}, a_{g^{-1}(3)}, a_{g^{-1}(4)})$

e.g., for g = (12)(34), g.H(w,x,y,z) = H(x,w,z,y)

- K-orbit associated to $(a, b, c, d) \in [\ell]^4$

$$\mathcal{O} = \{g \cdot (a, b, c, d) : g \in K\}$$

e.g., K-orbit of (a,a,b,c) is
{(a,a,b,c), (a,a,c,b), (b,c,a,a), (c,b,a,a)}

Theorem [R.'25]: Basis for K-Invariant Four-Variable Functions

For each K-orbit $\mathcal{O}$ on $[\ell]^4$, define the *orbit function*

$$\Psi_{\mathcal{O}} := \sum_{(a,b,c,d)\in\mathcal{O}} \varphi_a \otimes \varphi_b \otimes \varphi_c \otimes \varphi_d.$$

The collection of orbit functions

$$\mathcal{F} = \{\Psi_{\mathcal{O}} : \mathcal{O} \text{ is a } K\text{-orbit on } \{1,\ldots,\ell\}^4\}$$

forms an orthogonal basis for the K-invariant subspace U^K with respect to the inner product $\langle\cdot,\cdot\rangle_{\pi^{\otimes 4}}$.

$$\mathbb{E}[\varphi_a(W)|U] = e^{\lambda_a t_1}\varphi_a(U) \quad \text{(evolution from } r_1 \text{ to } s_1\text{)}$$

$$\mathbb{E}[\varphi_b(X)|U] = e^{\lambda_b t_2}\varphi_b(U) \quad \text{(evolution from } r_1 \text{ to } s_2\text{)}$$

$$\mathbb{E}[\varphi_c(Y)|V] = e^{\lambda_c t_3}\varphi_c(V) \quad \text{(evolution from } r_2 \text{ to } s_3\text{)}$$

$$\mathbb{E}[\varphi_d(W)|V] = e^{\lambda_d t_4}\varphi_d(V) \quad \text{(evolution from } r_2 \text{ to } s_4\text{)}$$

$$\begin{aligned}
&\mathbb{E}[\varphi_a(W)\varphi_b(X)\varphi_c(Y)\varphi_d(Z)] \\
&= \mathbb{E}[\mathbb{E}[\varphi_a(W)\varphi_b(X)\varphi_c(Y)\varphi_d(Z)|U,V]] \\
&= \mathbb{E}[\mathbb{E}[\varphi_a(W)|U]\mathbb{E}[\varphi_b(X)|U]\mathbb{E}[\varphi_c(Y)|V]\mathbb{E}[\varphi_d(W)|V]] \\
&= e^{\lambda_a t_1+\lambda_b t_2+\lambda_c t_3+\lambda_d t_4}\mathbb{E}[(\varphi_a\varphi_b)(U)\cdot(\varphi_c\varphi_d)(V)] \\
&= e^{\lambda_a t_1+\lambda_b t_2+\lambda_c t_3+\lambda_d t_4}\langle f_{ab}, e^{Qt_0} f_{cd}\rangle_\pi
\end{aligned}$$

Markov property

where $f_{ij} := \varphi_i\varphi_j$ pointwise

BONUS: Only depends on the orbit

Rate Matrix and Eigenbasis for Binary Model

For $\ell = 2$ with state space $\{1, 2\}$:

$$Q = \begin{pmatrix} -\pi_2 & \pi_2 \\ \pi_1 & -\pi_1 \end{pmatrix}$$

$$\varphi_1(1) = \varphi_1(2) = 1 \qquad \varphi_2(1) = \sqrt{\frac{\pi_2}{\pi_1}},\ \varphi_2(2) = -\sqrt{\frac{\pi_1}{\pi_2}}$$

$$\lambda_1 = 0, \quad \lambda_2 = -1$$

Theorem: Impossibility Result for GTR on $\ell = 2$ States

For any GTR model on $\ell = 2$ states, there exists no linear score.

Proof idea

Expectation must be zero on a mixture of stars trees *for any choice of pendant branch lengths*. Constrains all coefficients in the basis expansion to be zero.

This result also follows from [Matsen-Mossel-Steel'08] and, in the special case where π is uniform, from [Stefankovic-Vigoda'07] and [Matsen-Steel'07] via non-identifiability arguments.

Two Independent Binary Sites: Construction

Setup: Each taxon has two independent binary sites (effectively 4 states)

$$A \equiv (1,1), \quad B \equiv (1,2), \quad C \equiv (2,1), \quad D \equiv (2,2)$$

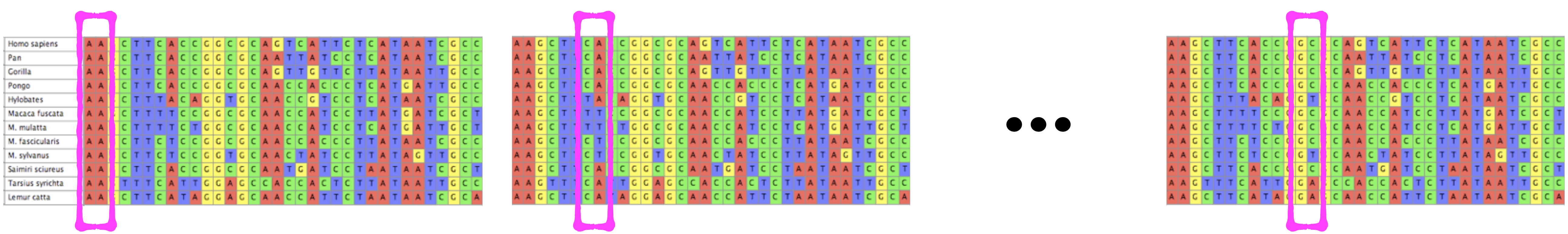
Rate matrix: sum of Kronecker products (reversible w.r.t. $\pi^{\otimes 2}$):

$$Q^{(2)} = Q \otimes I_2 + I_2 \otimes Q = \begin{pmatrix} -2\pi_2 & \pi_2 & \pi_2 & 0 \\ \pi_1 & -(\pi_1+\pi_2) & 0 & \pi_2 \\ \pi_1 & 0 & -(\pi_1+\pi_2) & \pi_2 \\ 0 & \pi_1 & \pi_1 & -2\pi_1 \end{pmatrix}$$

Eigenfunctions: Tensor products of single-site eigenfunctions:

$$\Phi_A = \varphi_1 \otimes \varphi_1 \quad \Phi_B = \varphi_1 \otimes \varphi_2 \quad \Phi_C = \varphi_2 \otimes \varphi_1 \quad \Phi_D = \varphi_2 \otimes \varphi_2$$

$$\Lambda_A = 0 \quad \Lambda_B = -1 \quad \Lambda_C = -1 \quad \Lambda_D = -2$$



Theorem: Linear Score for Two-Site Binary GTR Model [R.'25]

Let $\mathcal{O}_1$ be the K-orbit of (B, B, C, C) and $\mathcal{O}_2$ be the K-orbit of (B, C, B, C). Define $H = \Psi_{\mathcal{O}_1} - \frac{1}{2}\Psi_{\mathcal{O}_2}$. Then for $\vec{t} > 0$

$$\mathbb{E}_{T_3}[H] = 2e^{-\sum_{i=1}^{4} t_i}(1 - e^{-2t_0}) > 0$$

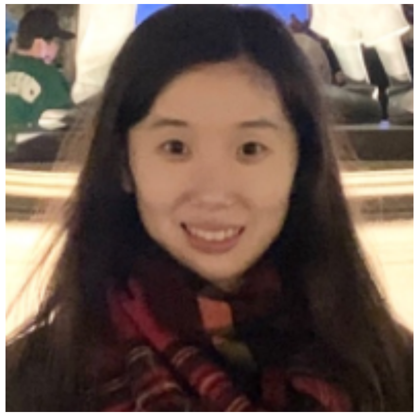
$$\mathbb{E}_{T_3}[(1\,4) \cdot H] = e^{-\sum_{i=1}^{4} t_i}(e^{-2t_0} - 1) < 0$$

$$\mathbb{E}_{T_3}[(2\,4) \cdot H] = e^{-\sum_{i=1}^{4} t_i}(e^{-2t_0} - 1) < 0.$$

Conclusion

A linear score exists for the *two-site* binary GTR model.

The special case π uniform was first studied in the Ph.D. thesis of former UW-Madison student Shuqi Yu.



IV. Generalizations

Notation: at $t_0 = 0$, the factor $\langle f_{ab}, e^{Qt_0} f_{cd} \rangle_\pi$ becomes

$$\langle \varphi_a \varphi_b, \varphi_c \varphi_d \rangle_\pi = \langle 1, \varphi_a \varphi_b \varphi_c \varphi_d \rangle_\pi =: K_{\{\{a,b,c,d\}\}}$$

Assumptions:

- (Λ): $\lambda_1 = 0 > \lambda_2 = -1 > \cdots > \lambda_\ell$ (i.e., eigenvalues of Q are distinct)
- (Φ): $K_{\{\{i,j,k,l\}\}} \neq 0$ for any multiset of four non-trivial (i.e., $\neq 1$) indices that are not all identical

Theorem: Impossibility Result [R.'25]

For any (single-site) GTR model on $\ell \geq 2$ states, if (Λ) and (Φ) hold, then there exists no linear score.

Assumptions:

- (Φ): $K_{\{\{i,j,k,l\}\}} = \langle 1, \varphi_i \varphi_j \varphi_k \varphi_l \rangle_\pi \neq 0$ for any multiset of four non-trivial (i.e., $\neq 1$) indices that are not all identical

Not necessary

Possible for (Φ) to fail, yet no linear score exists (e.g., K3P [R.'25]).

Theorem: Disjoint Support Trick [R.'25]

For any (single-site) GTR model on $\ell \geq 2$ states where (Φ) fails because two eigenfunctions φ_a, φ_b have disjoint support, there exists a linear score.

Proof idea

Let $\mathcal{O}$ be K-orbit of (a, a, b, b) and $H = \Psi_{\mathcal{O}}$. Positive on T_3, 0 on T_2, T_1 (so linear invariant; e.g., TN93 case [Casanellas-Homs-Torres'24]).

Assumptions:

- (Λ): $\lambda_1 = 0 > \lambda_2 = -1 > \cdots > \lambda_\ell$ (i.e., eigenvalues of Q are distinct)

Theorem: Distinct Eigenvalues Trick [R.'25]

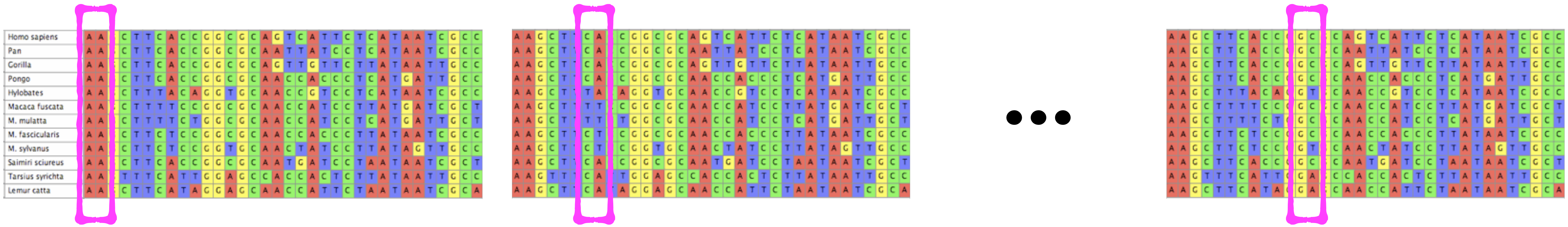
For any (single-site) GTR model on $\ell \geq 2$ states where (Λ) fails, there exists a linear score.

Proof idea

Assume $\lambda_a = \lambda_b$. Let $\mathcal{O}_1$ be the K-orbit of (a, a, b, b) and $\mathcal{O}_2$ be the K-orbit of (a, b, a, b). Define $H = \Psi_{\mathcal{O}_1} - \frac{1}{2}\Psi_{\mathcal{O}_2}$.

Two-site setting

- States: $(a, b) \in [\ell]^2$ numbered lexicographically
- Rate matrix: $Q^{(2)} = Q \otimes I_2 + I_2 \otimes Q$
- Eigenfunctions: $\varphi^{(2)}_{(a,b)} = \varphi_a \otimes \varphi_b$ with eigenvalue $\lambda^{(2)}_{(a,b)} = \lambda_a + \lambda_b$



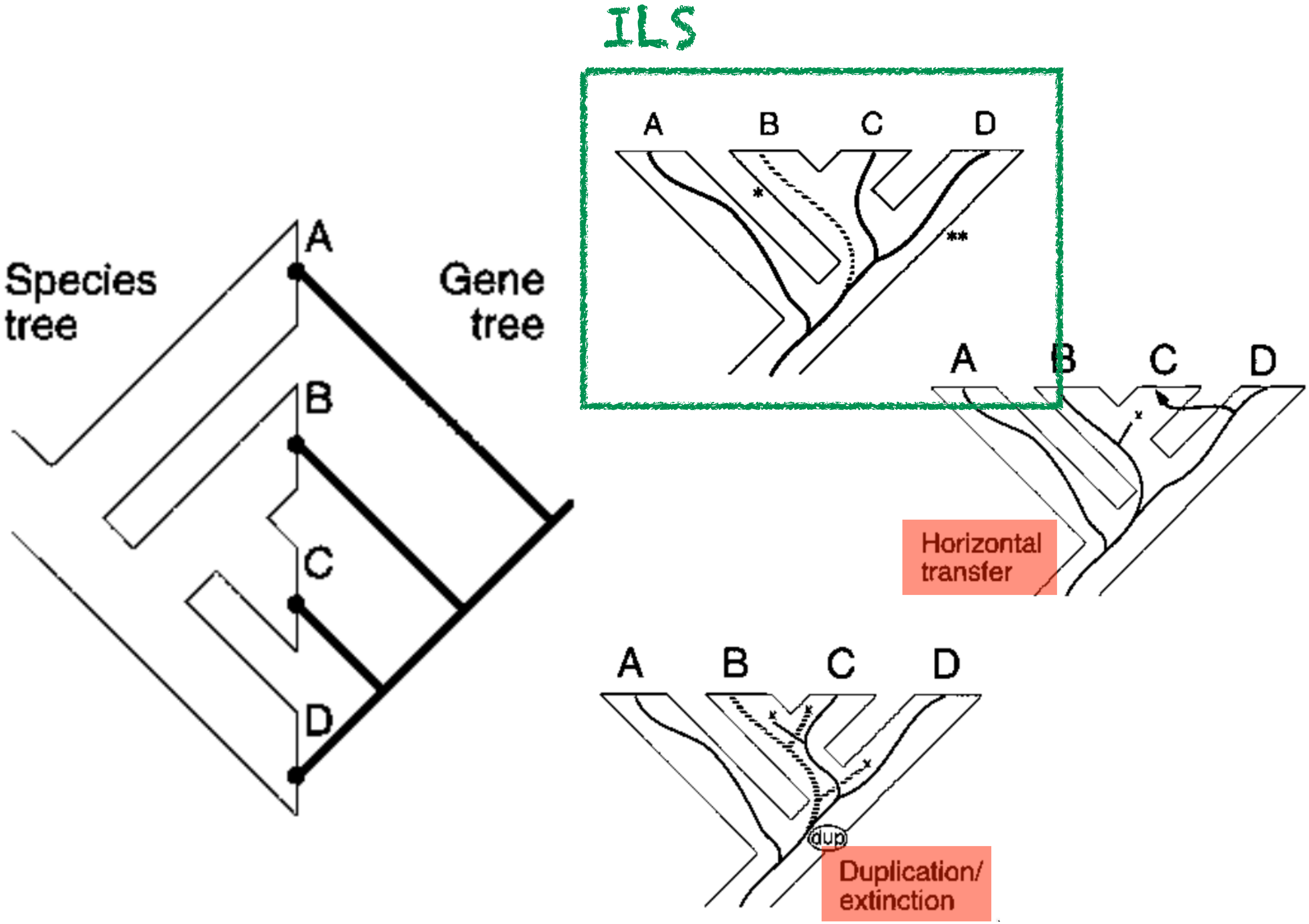
Theorem: The Power of Pairs of Sites [R.'25]

For any two-site GTR model on $\ell \geq 2$ states, there exists a linear score.

Proof idea

For any $a \neq b$,

$$\lambda^{(2)}_{(a,b)} = \lambda_a + \lambda_b = \lambda_b + \lambda_a = \lambda^{(2)}_{(b,a)}$$

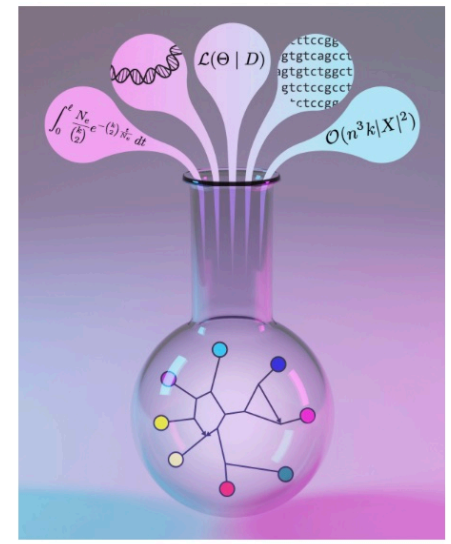


Figures from Maddison, Syst. Biol. (1997).

Institute for Mathematical and Statistical Innovation

Theory, Methods, and Applications of Quantitative Phylogenomics

ICERM



Sep 4 - Dec 6, 2024
Semester Program

Thank you for your attention

Work supported by:

SIMONS FOUNDATION