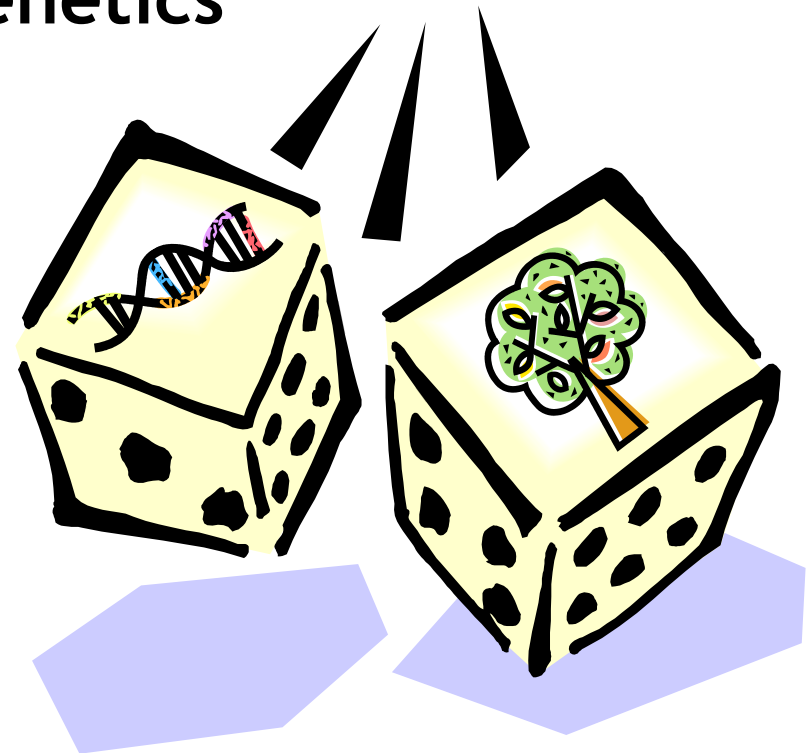


Probabilistic Techniques in Mathematical Phylogenetics

Sebastien Roch
UCLA

SoCal Probability Symposium
Dec 5, 2009



outline of the talk

PART I

background: phylogenetic reconstruction



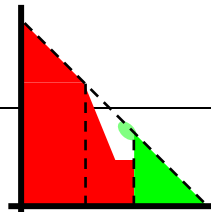
PART II

the power of the distance matrix



PART III

reconstruction on trees:
exponential moment

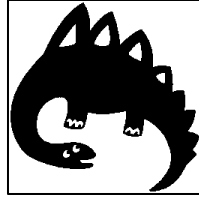
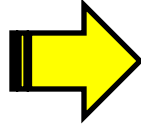


PART I

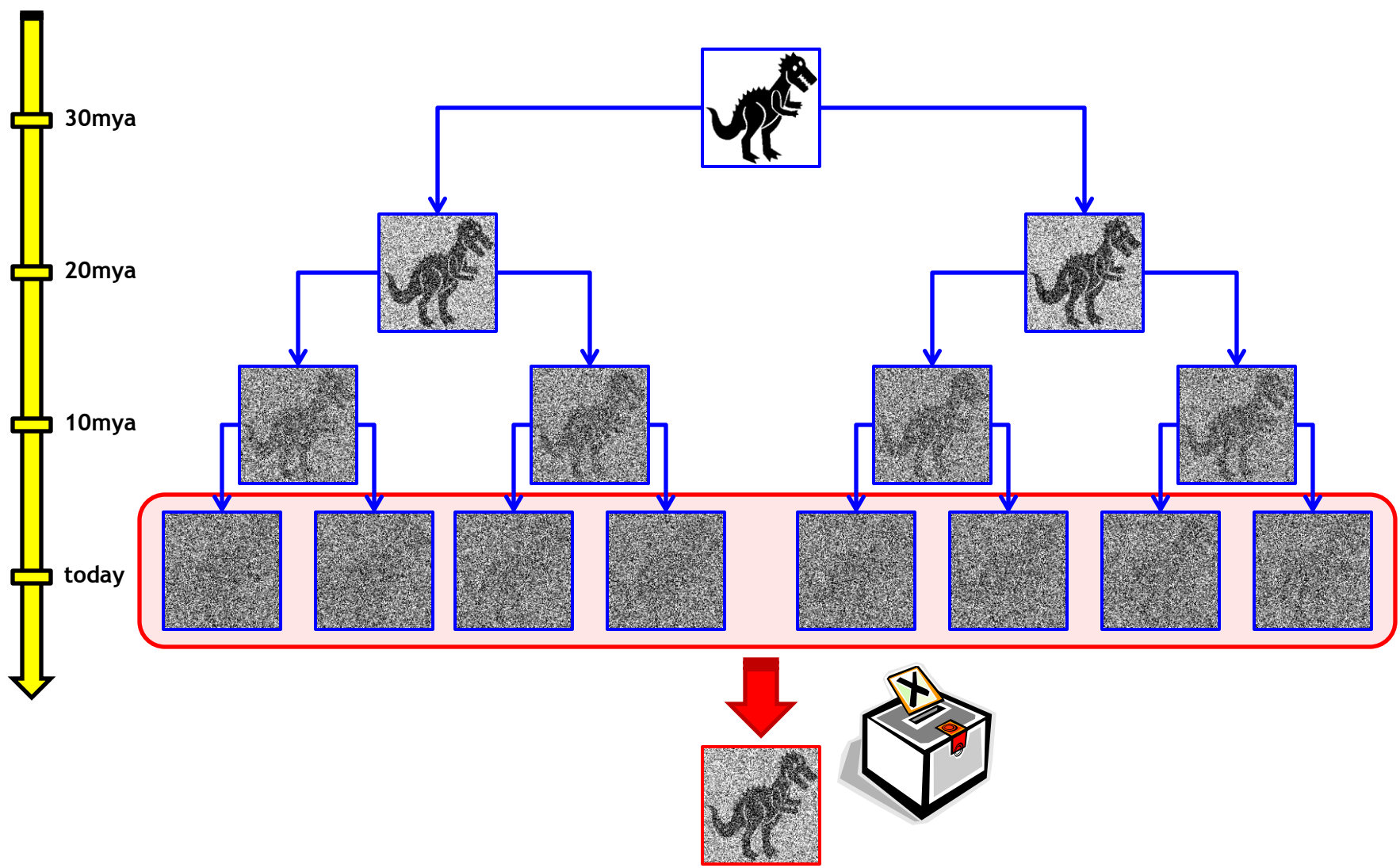
background:

phylogenetic reconstruction









Markov chain on a tree

- **broadcasting model**

- b-ary tree: $T = (V, E)$
- node states:

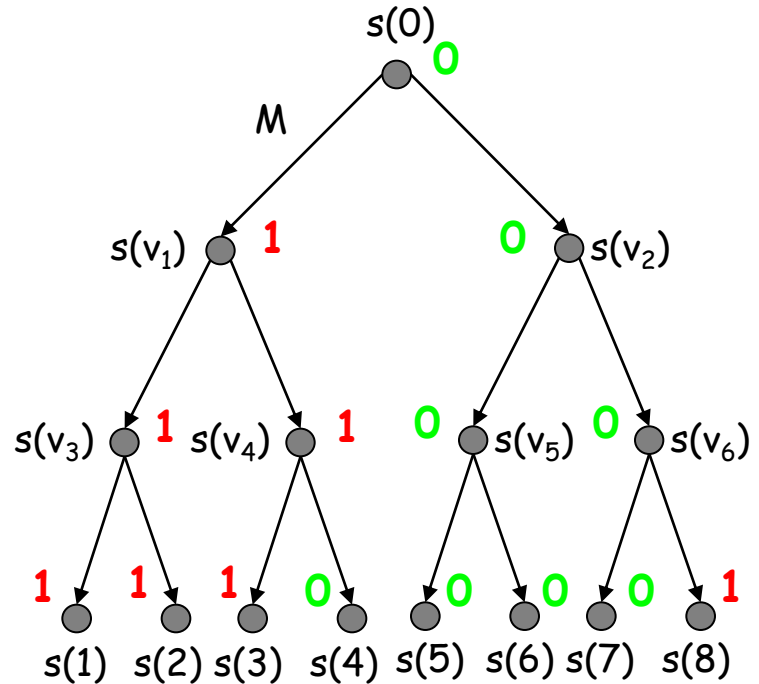
$$\{s(v) \in \{0,1\} : v \in V\}$$

- Markov transition matrix:

$$M = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

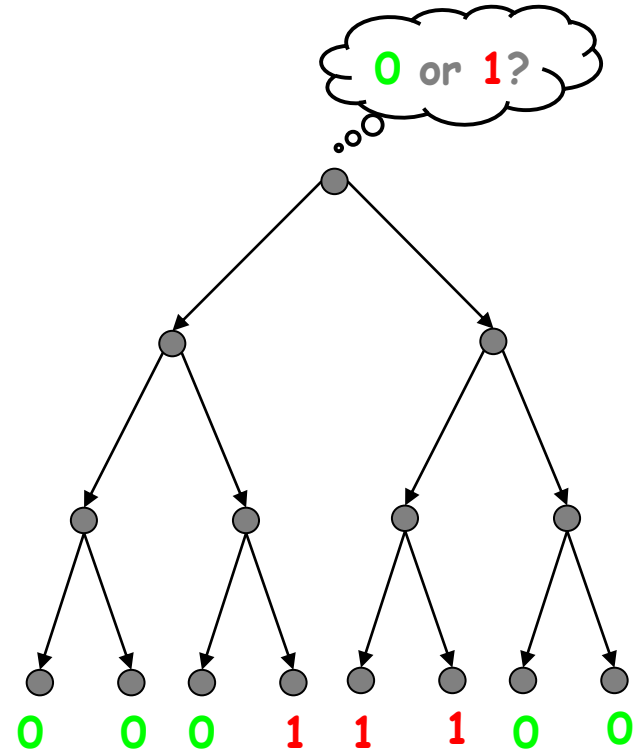
- stationary distribution:

$$\pi = (\pi_0, \pi_1)$$



the reconstruction problem

- ancestral reconstruction
 - **given**: states at leaves
 - **goal**: infer state at root
- phase transition
 - trade-off between **noise** and **duplication**



inferring ancestral states I

- **majority** estimator [KS67,K77,MS79]

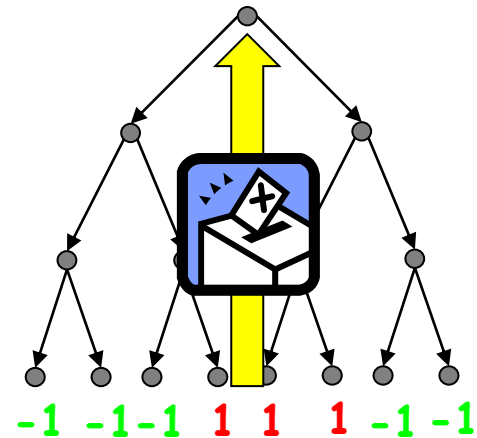
$$Z_h = (bg)^{-h} \sum_{i \in [n]} s_i$$

where $g=1-2p$, $h=\text{\#levels}$, $n=\text{\#leaves}$

- **properties:** conditional expectation

$$\begin{aligned} E_+^h[s_1] &= (1-p) E_+^{h-1}[s_1] + p E_-^{h-1}[s_1] \\ &= (1-2p) E_+^{h-1}[s_1] = \dots = g^h \end{aligned}$$

$$\Rightarrow E_{s_r}^h[Z_h] = s_r \quad \& \quad E^h[Z_h] = 0$$



- **properties:** variance

$$1) \text{Var}^h[Z_h] = \text{E}^h[Z_h^2] = \text{E}_+^h[Z_h^2] = \text{E}_-^h[Z_h^2]$$

$$2) \text{Var}^h[Z_h] = \text{Var}^h[s_r] + \text{Var}_+^h[Z_h]$$

$$= 1 + b \text{Var}_+^h[Z_h^{(1)}] \quad \text{(conditional independence)}$$

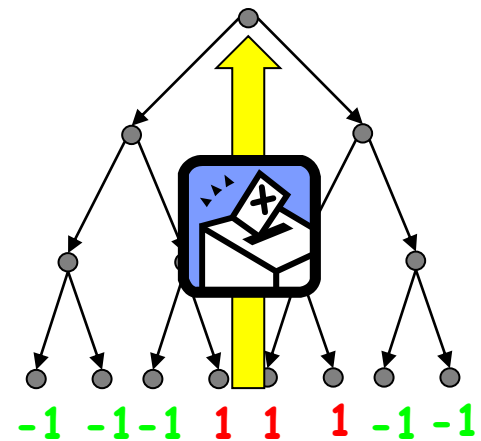
$$= 1 + b \text{E}_+^h[(Z_h^{(1)})^2] - b(\text{E}_+^h[Z_h^{(1)}])^2$$

$$= (1 - b^{-1}) + b \left\{ (1 - p)(bg)^{-2} \text{E}_+^{h-1}[(Z_{h-1})^2] + p(bg)^{-2} \text{E}_-^{h-1}[(Z_{h-1})^2] \right\} \quad \text{(Markov)}$$

$$= (1 - b^{-1}) + (bg^2)^{-1} \text{Var}^{h-1}[Z_{h-1}]$$

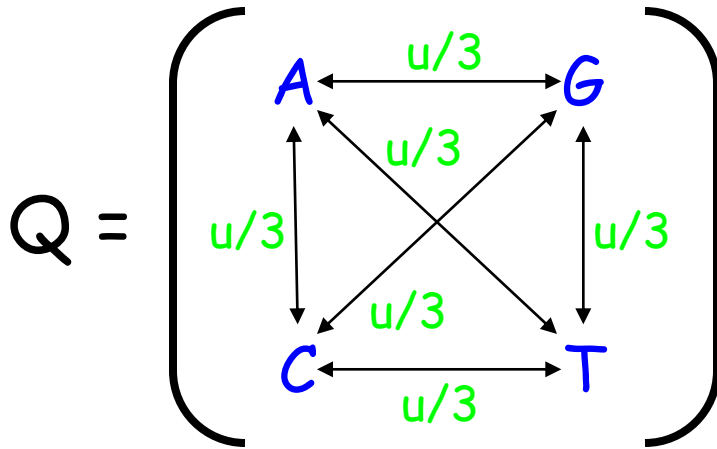
$$\rightarrow \begin{cases} +\infty & bg^2 < 1 \\ \frac{1 - b^{-1}}{1 - (bg^2)^{-1}} & bg^2 > 1 \end{cases}$$

best possible
for 2-state symmetric case
[BRZ95]

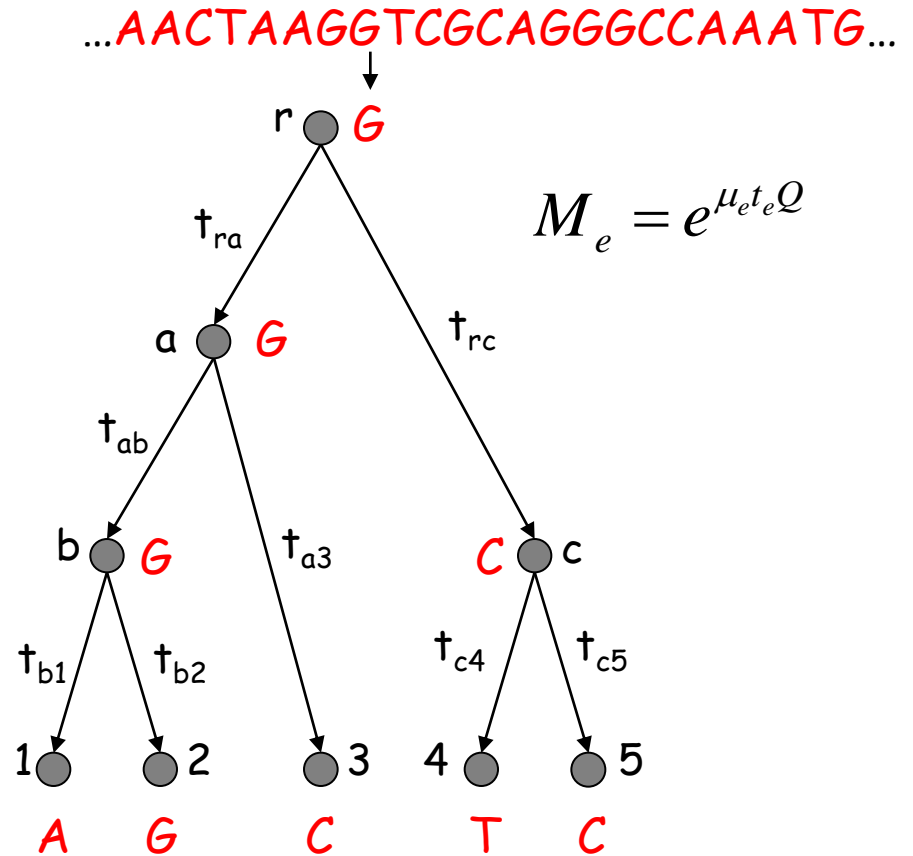


GTR model of evolution

- example: Jukes-Cantor model/Potts model with free boundary
 - phylogeny: T
 - number of species: n
 - number of states: q (=4)

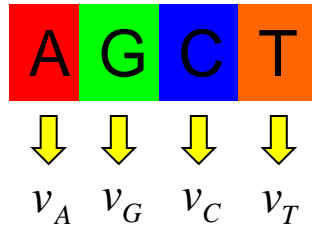


- note
 - no deletion/insertion



linear ancestral estimators

- **Kesten-Stigum bound for GTR** [KS67, MP03]
 - root estimator: v second eigenvector of Q

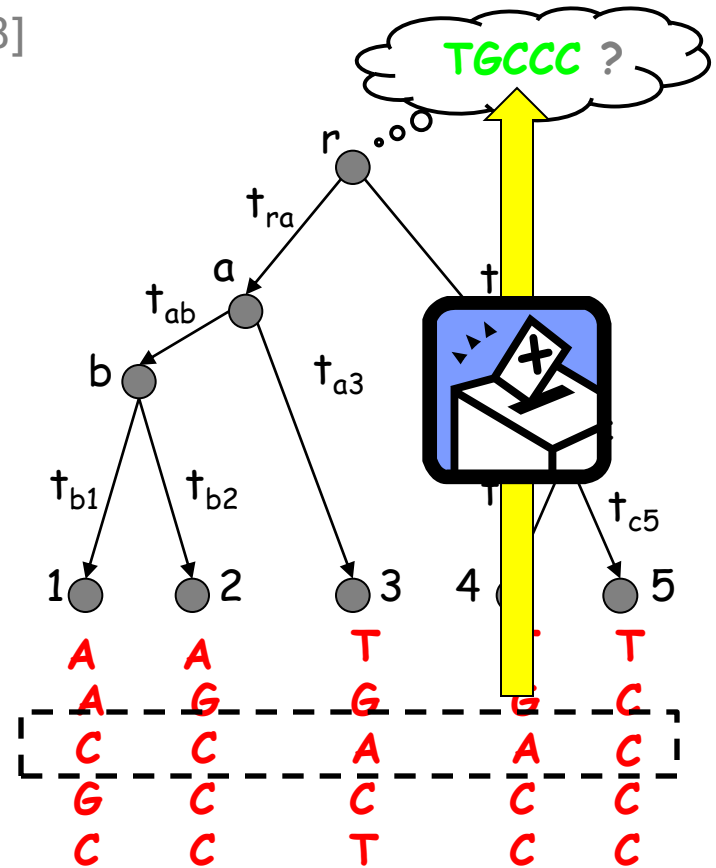


$$Z_{[n]} = \sum_{x \in [n]} 2^{-|x|} e^{\mu_{rx} t_{rx}} s_x^i$$

- critical branch length:

$$G^* = \ln \sqrt{2}$$

- conditionally unbiased and **bounded variance** below G^*



asymptotic sample complexity

- **setup**

- trees on n leaves: \mathcal{T}_n
- model: $(\mathcal{T}, \{t_e\}_{e \in E})$ in Θ_n
- k i.i.d. samples: s_L^1, \dots, s_L^k
- estimator:

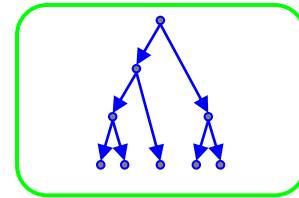
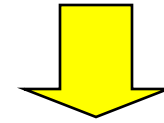
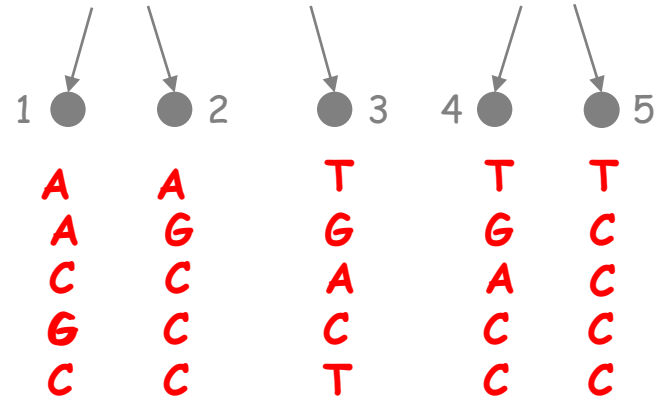
$$\Psi_n : \left\{ s_L^i \right\}_{i=1}^k \mapsto T \in \mathcal{T}_n$$

- **definition** - the estimator Ψ_n solves the **phylogenetic reconstruction problem** with k samples and confidence $1-\delta$ if for all models $(\mathcal{T}, \{t_e\}_{e \in E})$ in Θ_n

$$\mathbb{P} \left[\Psi_n \left(\left\{ s_L^i \right\}_{i=1}^k \right) = T \right] \geq 1 - \delta$$

- **convergence**

- how does k scale as a function of n ?





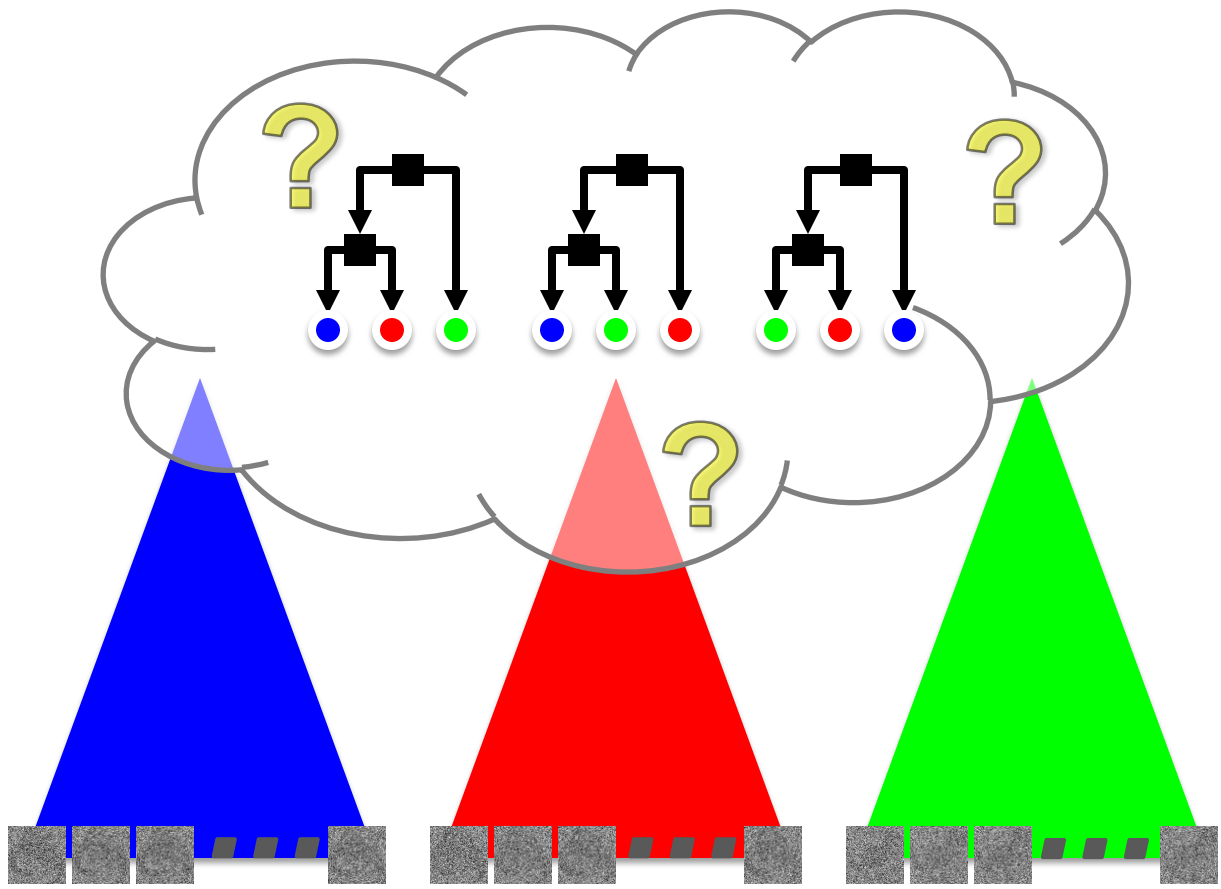
40mya

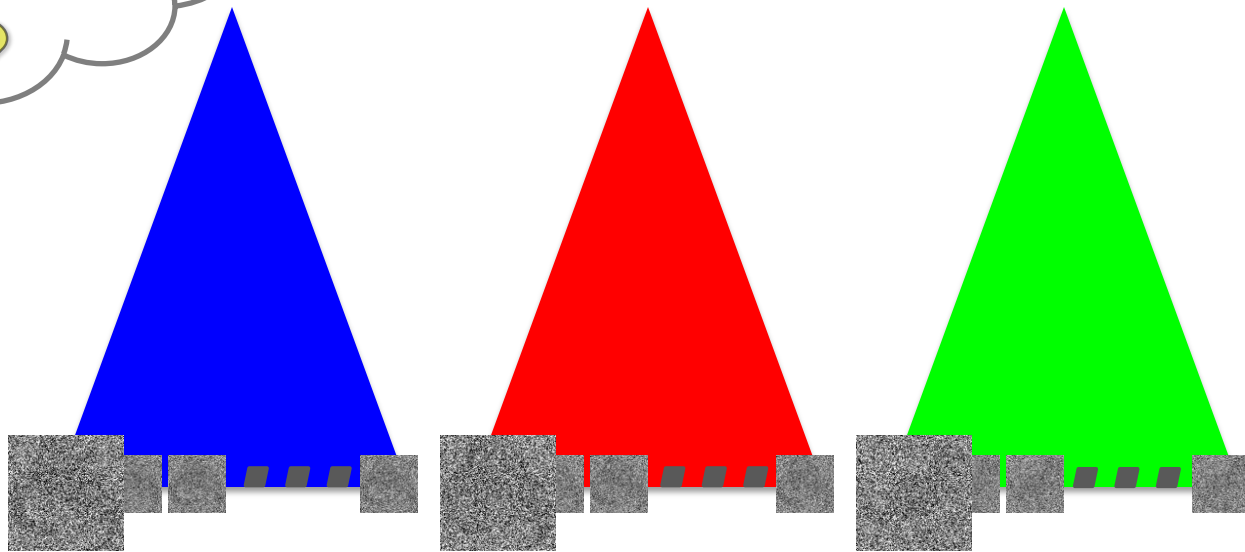
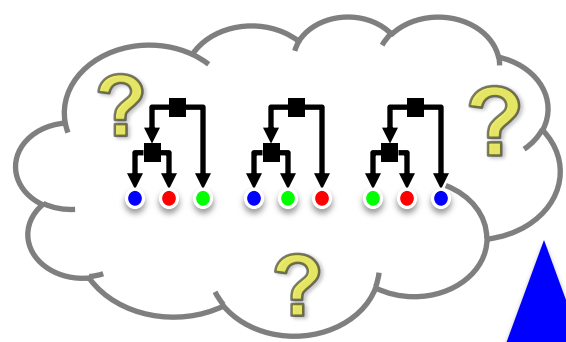
30mya

20mya

10mya

today



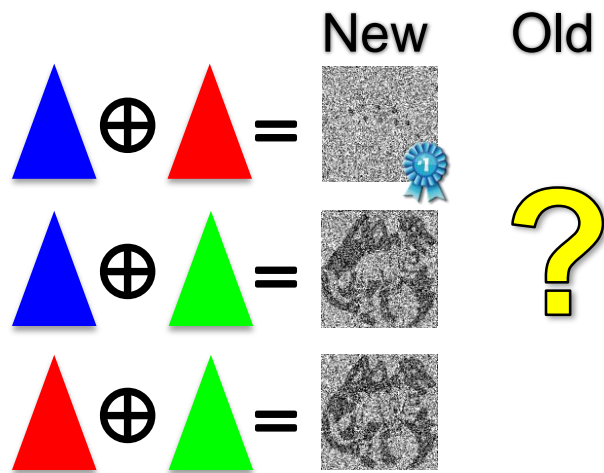
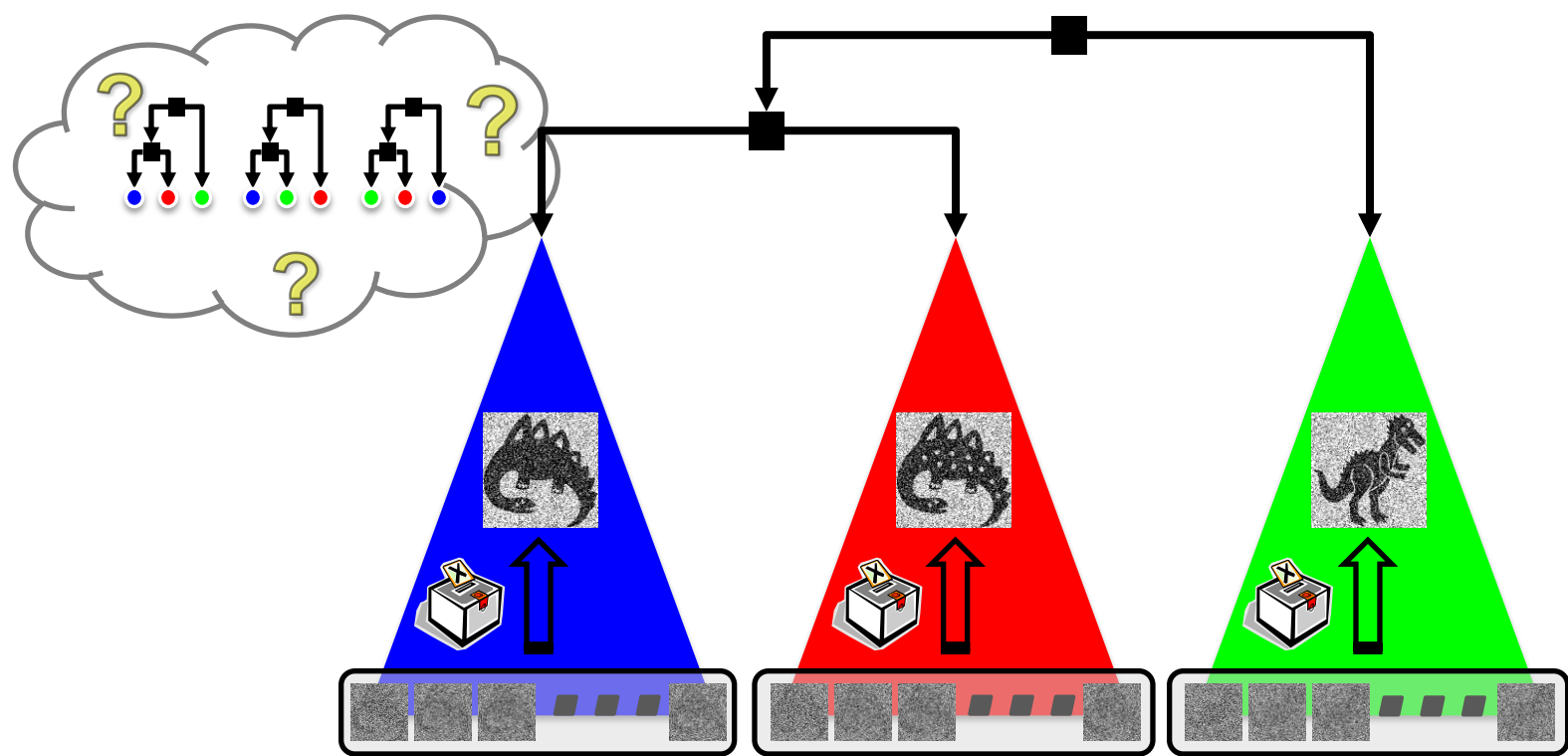


$$\triangle_{\text{blue}} \oplus \triangle_{\text{red}} = \text{gray block}$$

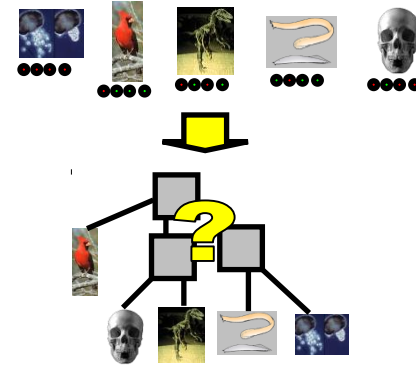
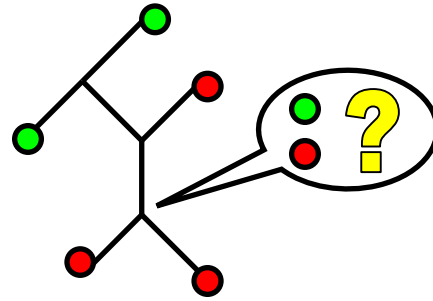
$$\triangle_{\text{blue}} \oplus \triangle_{\text{green}} = \text{gray block}$$

$$\triangle_{\text{red}} \oplus \triangle_{\text{green}} = \text{gray block}$$

?



Steel's conjecture



ancestral reconstruction		phylogenetic reconstruction
reconstruction		seq. length = $c \log n$
non-reconstruction		seq. length = n^c

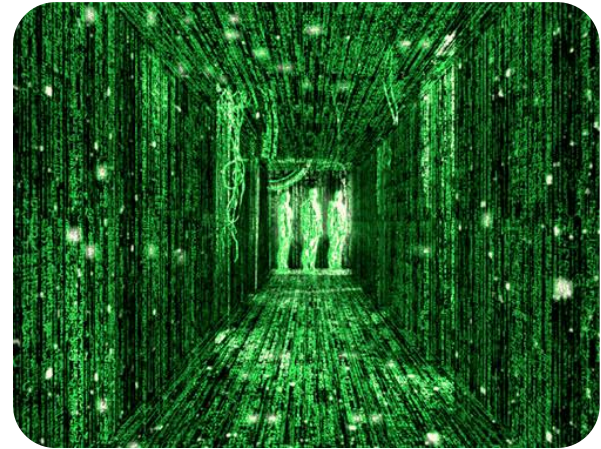
[Daskalakis-Mossel-R'06]

[Mossel'04]

$n = \# \text{ species}$

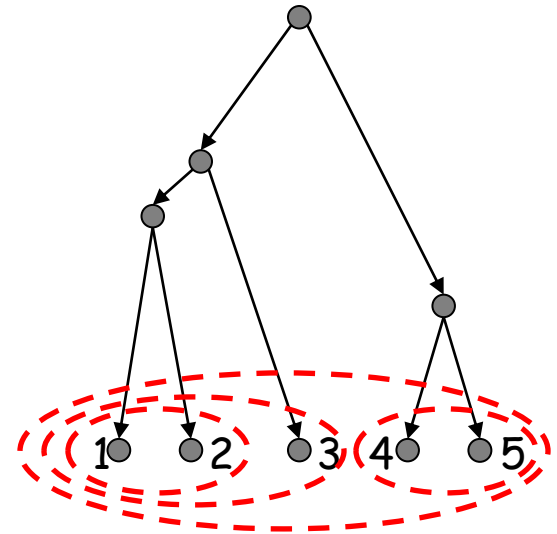
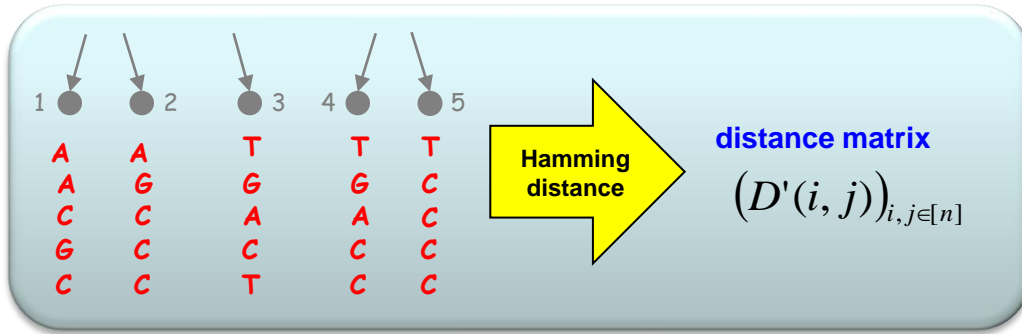
PART II

tree building methods:
the power of the distance matrix



- **assumption (A)** - assume $0 < F < u(e)t(e) < G < G^*$, for all e
- **theorem [R.]** - under (A) + discretization, there is a distance-based method that only requires

$$k \propto O(\log n)$$

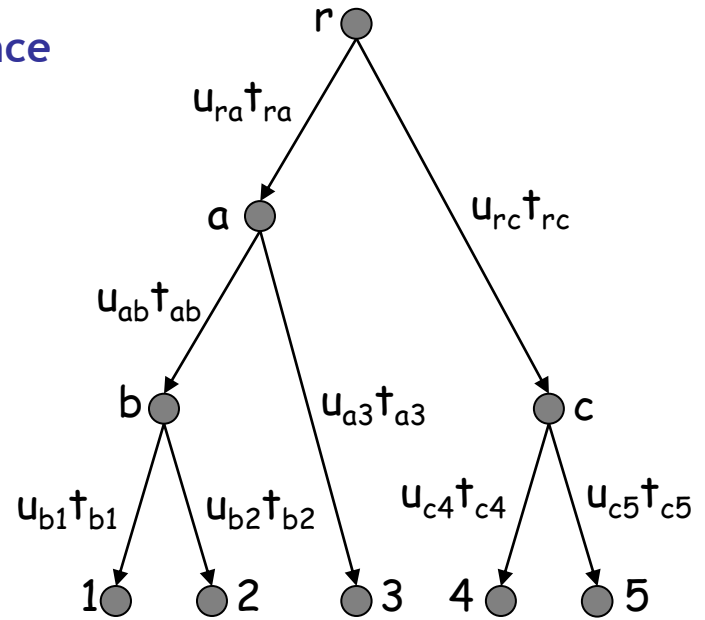


distance-matrix methods

- in our case:
 - associate to each pair of leaves a **distance**

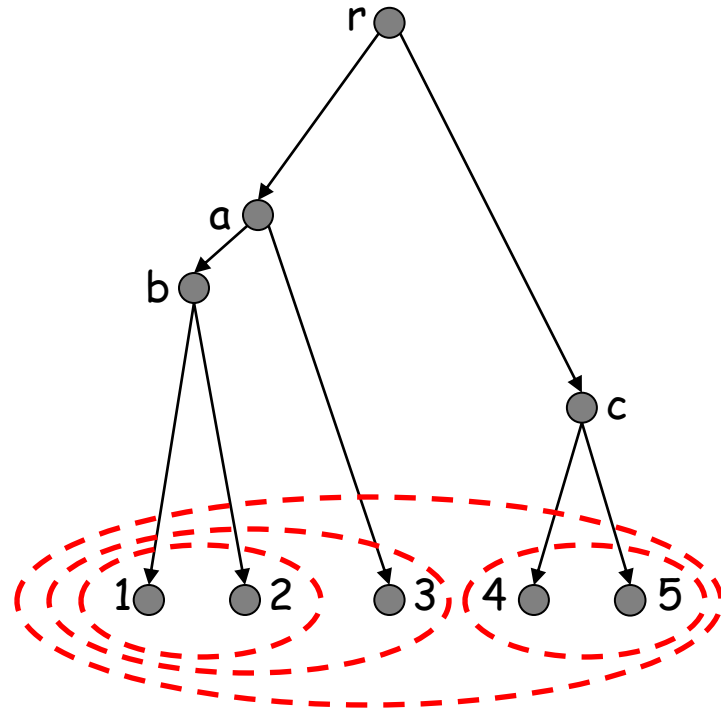
$$D(i, j) = \sum_{e \in P(T; i, j)} \mu_e t_e$$

- defines a **tree metric**
- key property:
 - completely characterizes the tree
- reconstruction algorithm:
 - estimate $D(i, j)$ from sequences
 - deduce the topology of the tree
- **fact** - reconstruction can be done very efficiently
 - e.g. UPGMA, Neighbor Joining (NJ), Short Quartet Method (SQM)



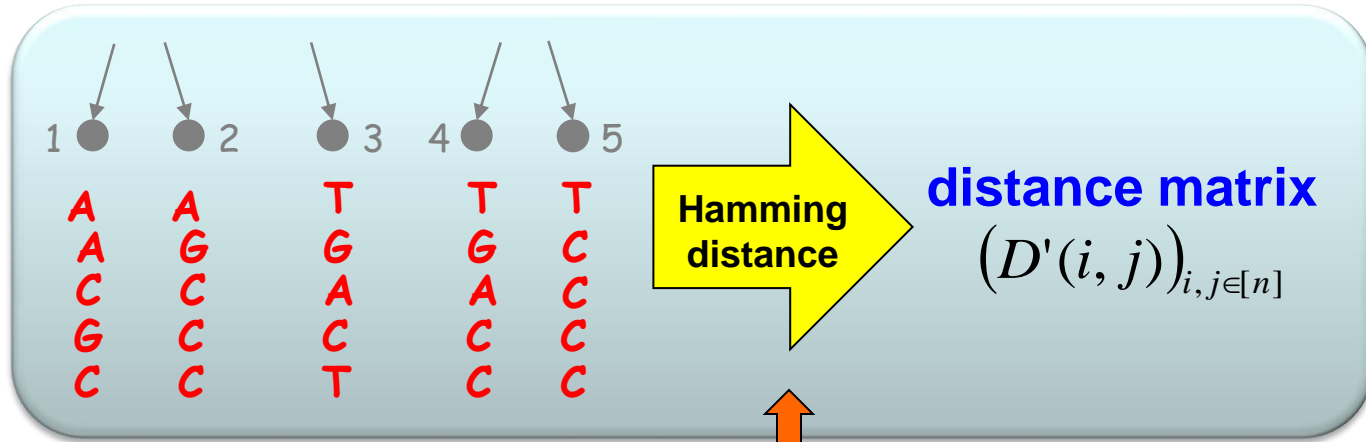
distance estimates: $(D'(i, j))_{i, j \in [n]}$

$\mu = 1$



- **assumption (B)** - assume $0 < F < t(e) < G$, for all e
- **theorem [ESSW'99]** - under (B), UPGMA and Short Quartet Method need polynomial-length sequences, i.e.,

$$k \propto n^c$$



$$\frac{H(s_i, s_j)}{k} = \frac{3}{4} (1 - e^{-D'(i, j)})$$

loss of information?

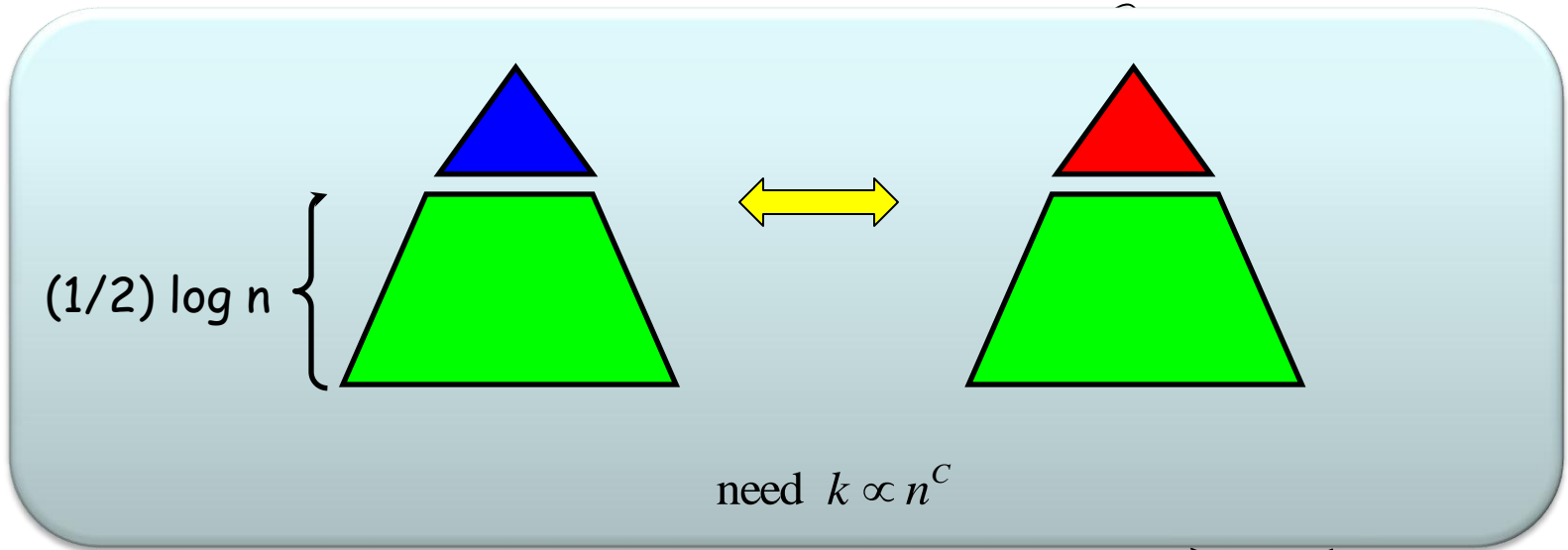
$$j) = O(\log n)$$

- **fact** - to estimate distances of order M with precision ε , one needs

$$k \propto \frac{e^M}{\min\{1, \varepsilon^2\}} \log n$$

- **definition** [King et al.'03, Mossel'07] - a symmetric matrix D' is a (ε, M) -**distortion** of the distance matrix D if

$$|D'(i, j) - D(i, j)| < \varepsilon \text{ if } D'(i, j) < M + \varepsilon \text{ or } D(i, j) < M + \varepsilon$$



- **observation** - the entries of the distance matrix are correlated random variables. in particular, the joint distribution of

$$(D'(a,b), D'(c,d))$$

depends on the joint distribution of states at (a,b,c,d)

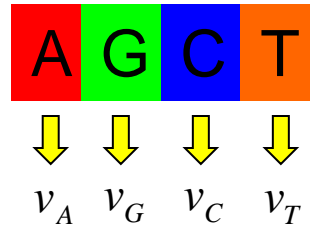
$$\mu_{\{a,b,c,d\}}$$

- **questions**
 - how to extract this extra information?
 - how useful is it really?



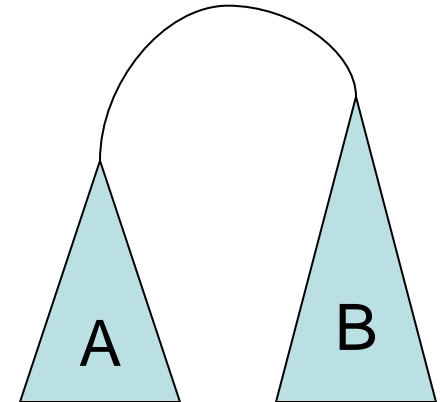
revisiting the averaging procedure I

- **step 1** - project the states to second eigenvector



- the distance matrix becomes

$$D'(a, b) = -\ln\left(\frac{1}{k} \sum_{i=1}^k s_a^i s_b^i\right)$$



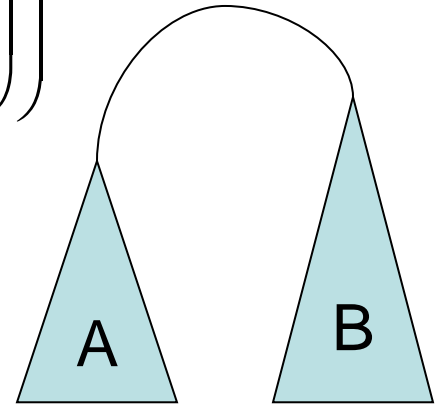
revisiting the averaging procedure II

- **step 2** - perform “exponential averaging” between clusters

$$\begin{aligned}
 D((A,B)) &= -\ln \left(\frac{1}{|A|+|B|} \sum_{a \in A} \sum_{b \in B} 2^{-|a|-|b|} e^{-D'(a,b)} \right) \\
 &= -\ln \left(\sum_{a \in A} \sum_{b \in B} 2^{-|a|-|b|} \frac{1}{k} \sum_{i=1}^k s_a^i s_b^i \right) \\
 &= -\ln \left(\frac{1}{k} \sum_{i=1}^k \left(\sum_{a \in A} 2^{-|a|} s_a^i \right) \left(\sum_{b \in B} 2^{-|b|} s_b^i \right) \right)
 \end{aligned}$$

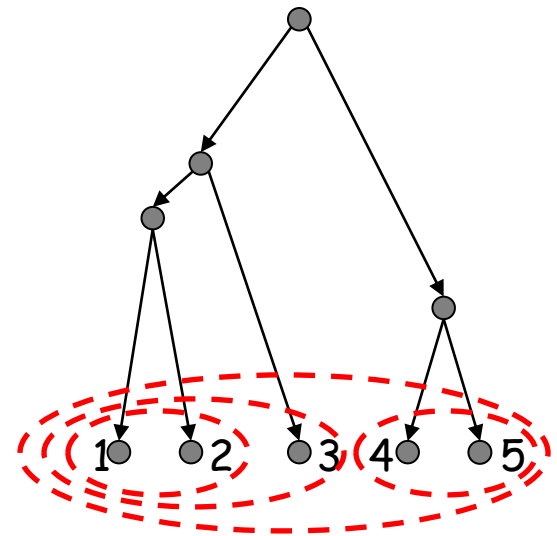
“majority”

$$D'(a,b) = -\ln \left(\frac{1}{k} \sum_{i=1}^k s_a^i s_b^i \right)$$



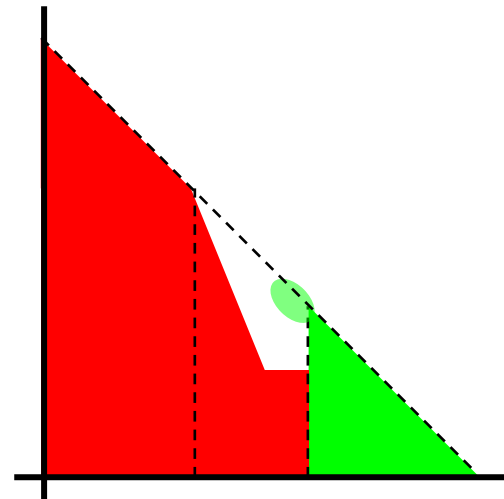
- **assumption (A)** - assume $0 < F < t(e) < G < G^*$, for all e
- **theorem [R.]** - under (A), WPGMA only requires

$$k \propto O(\log n)$$



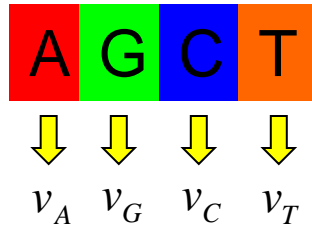
PART III

proof: exponential moment
of linear estimators



linear ancestral estimators

- **Kesten-Stigum bound for GTR** [KS67, MP03]
 - root estimator: v second eigenvector of Q

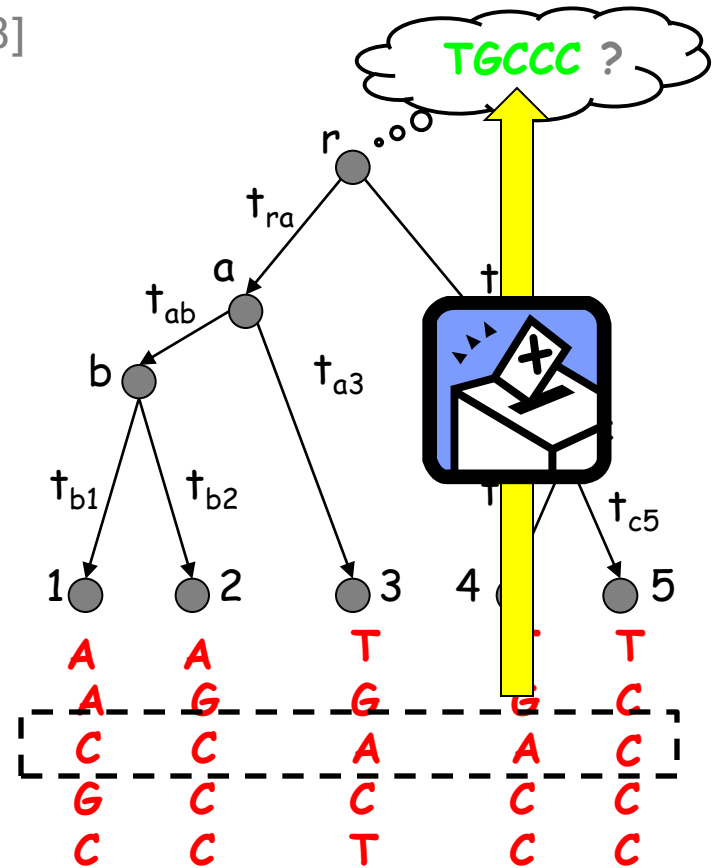


$$Z_{[n]} = \sum_{x \in [n]} 2^{-|x|} e^{\mu_{rx} t_{rx}} s_x^i$$

- critical branch length:

$$G^* = \ln \sqrt{2}$$

- conditionally unbiased and **bounded variance** below G^*



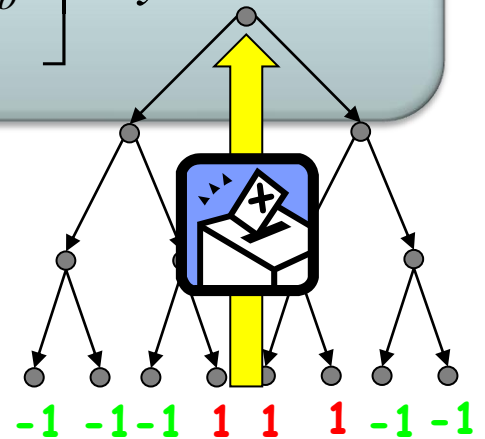
- **thm [Peres-R.]:** there exists $c'' > 0$ and $x'' > 0$ s.t. for all $h > 0$, x in $(-x'', x'')$ and i in $[q]$

$$E_i^h [\exp(xZ_h^2)] \leq c''$$

below t^*

- **corollary:** for all $x > 0$, there is $0 < y < 1$ s.t.

$$P \left[\left| \frac{1}{k} \sum_{i=1}^k Z_A^i Z_B^i - \frac{1}{k} \sum_{i=1}^k s_{a^*}^i s_{b^*}^i \right| > x \mid s_{a^*}, s_{b^*} \right] < y^k$$



- **thm [Peres-R.]:** there exists $c'' > 0$ and $x'' > 0$ s.t. for all $h > 0$, x in $(-x'', x'')$ and i in $[q]$

$$E_i^h[\exp(xZ_h^2)] \leq c'' < +\infty$$

below t^*

- **proof sketch of thm:** it suffices to prove that

$$\ln E_i^h[\exp(xZ_h)] \leq x E_i^h[Z_h] + cx^2$$

indeed, let N be an independent standard normal. applying Fubini twice

$$\begin{aligned} E_i^h[\exp(xZ_h^2)] &= E_i^h \left[E_N \left[\exp(\sqrt{2x}Z_h N) \right] \right] \\ &\leq E_N \left[\exp(v_i \sqrt{2x}N + c2xN^2) \right] \\ &< +\infty \end{aligned}$$

for x small

- **claim 1:** there exists $c > 0$ s.t. for all x in \mathbb{R} and i in $[q]$

$$\ln E_i^h[\exp(xZ_h)] \leq x E_i^h[Z_h] + cx^2$$

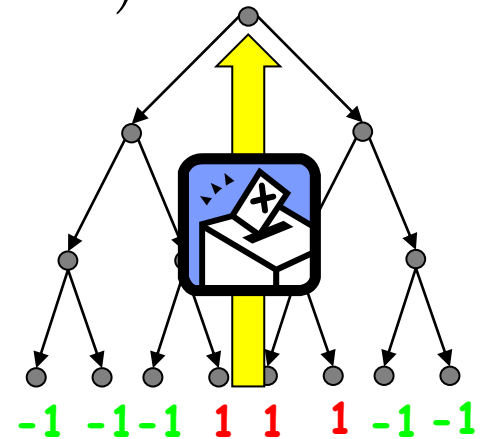
below t^*

- **proof sketch of claim 1:** let v be the second eigenvector of Q with corresponding eigenvalue -1 . we need the following fact

claim 2: for all $t > 0$ there exists $c' > 0$ s.t.

$$\sum_{j \in [q]} [e^{tQ}]_{ij} \exp(xv_j) \leq \exp(xv_i e^{-t} + c' x^2)$$

for all x in \mathbb{R} and i in $[q]$

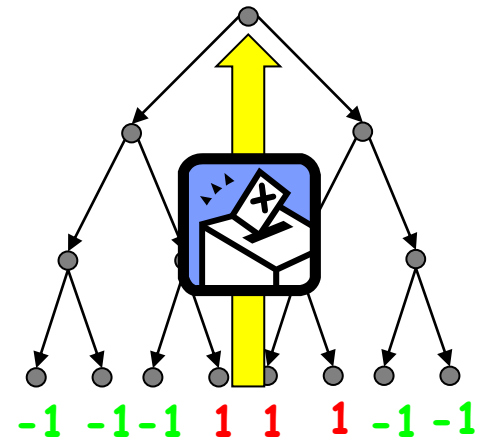


- claim 1: there exists $c > 0$ s.t. for all x in \mathbb{R} and i in $[q]$

$$\ln E_i^h[\exp(xZ_h)] \leq x E_i^h[Z_h] + cx^2$$

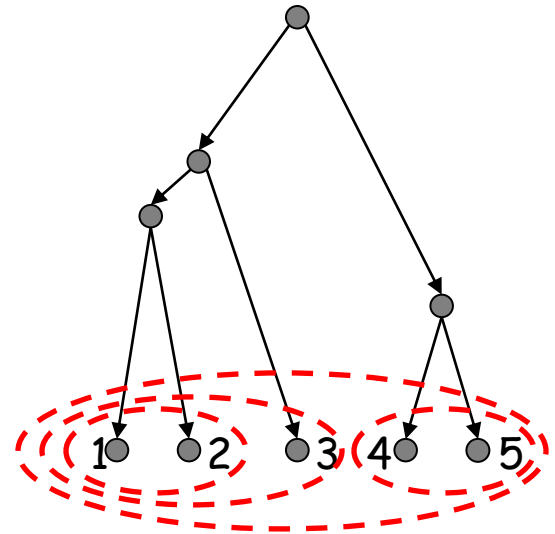
below t^*

$$\begin{aligned} & \ln E_i^h[\exp(x(Z_h^{(1)} + Z_h^{(2)}))] \\ &= 2 \ln E_i^h[\exp(xZ_h^{(1)})] \quad \text{(c.i.)} \\ &= 2 \ln \sum_{j \in [q]} [e^{tQ}]_{ij} E_j^{h-1}[\exp(x(2e^{-t})^{-1} Z_{h-1})] \quad \text{(M.)} \\ &= 2 \ln \sum_{j \in [q]} [e^{tQ}]_{ij} E_j^{h-1}[\exp(x(2e^{-t})^{-1} v_j + cx^2(2e^{-t})^{-2})] \\ &= 2 \{ cx^2(2e^{-t})^{-2} + x(2e^{-t})^{-1} v_i e^{-t} + c' x^2(2e^{-t})^{-2} \} \\ &= x v_i e^{-t} + (2e^{-2t})^{-1} (c + c') x^2 \end{aligned}$$



- **assumption (B)** - assume $0 < F < t(e) < G < t^*$, for all e
- **theorem [R.]** - under (B), WPGMA only requires

$$k \propto O(\log n)$$

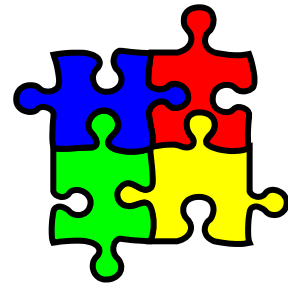
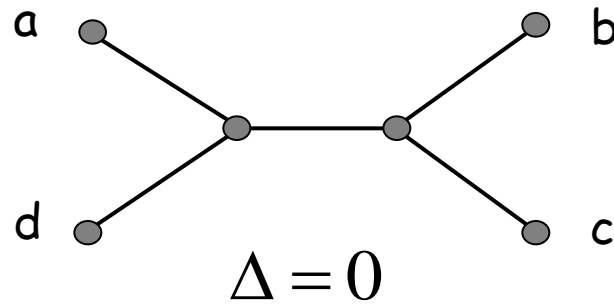
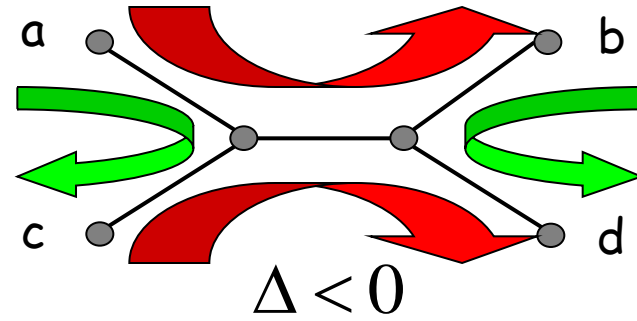
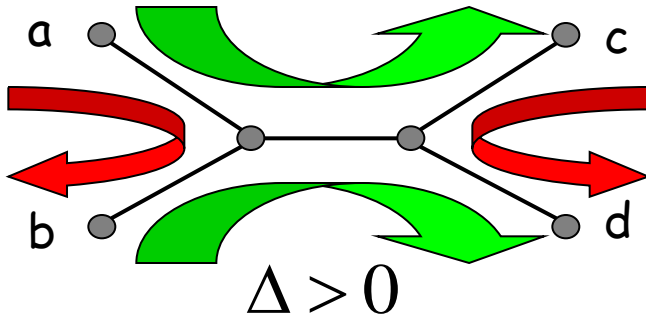


- **future directions**
 - sample complexity of MLE and parsimony
 - more realistic models (rates-across-sites, insertions/deletions)
 - non-discretized branch lengths



thank
you

$$\Delta = D'(a,c) + D'(b,d) - D'(a,b) - D'(c,d)$$



reconstruction algorithm

- loop
 - 1) distance estimation
 - 2) reconstruct one level
 - 3) infer sequences at roots

