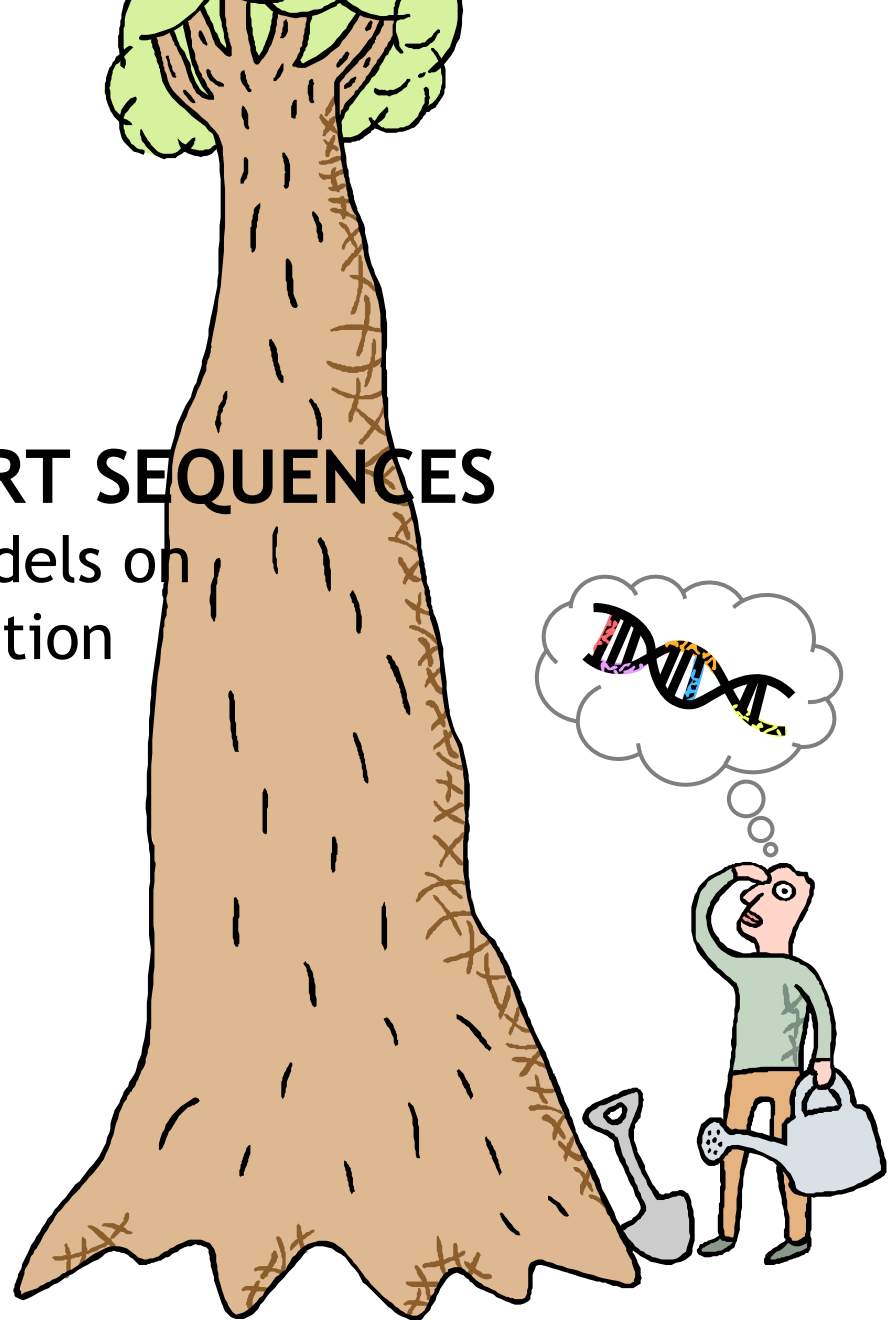


BIG TREES FROM SHORT SEQUENCES

reconstructing Markov models on
trees beyond the KS transition

Sebastien Roch
Microsoft Research

joint work with:
Elchanan Mossel, Allan Sly



outline of the talk

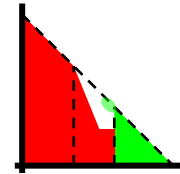
PART 0
review: statistical
phylogenetics



PART I
insights from
statistical physics

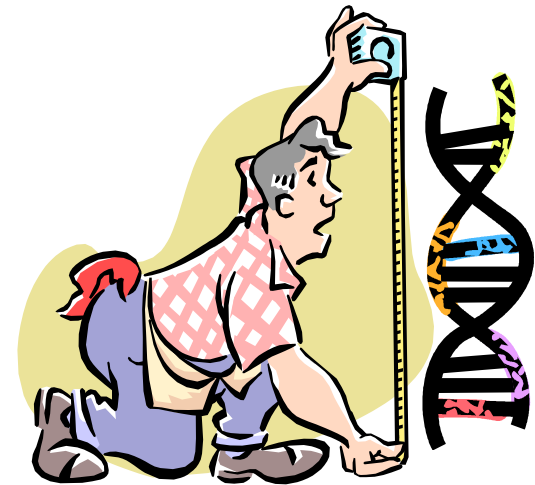


PART II
beyond the
KS transition



PART 0

background: phylogenetics



stochastic model of evolution

- Jukes-Cantor model/Potts model with free boundary

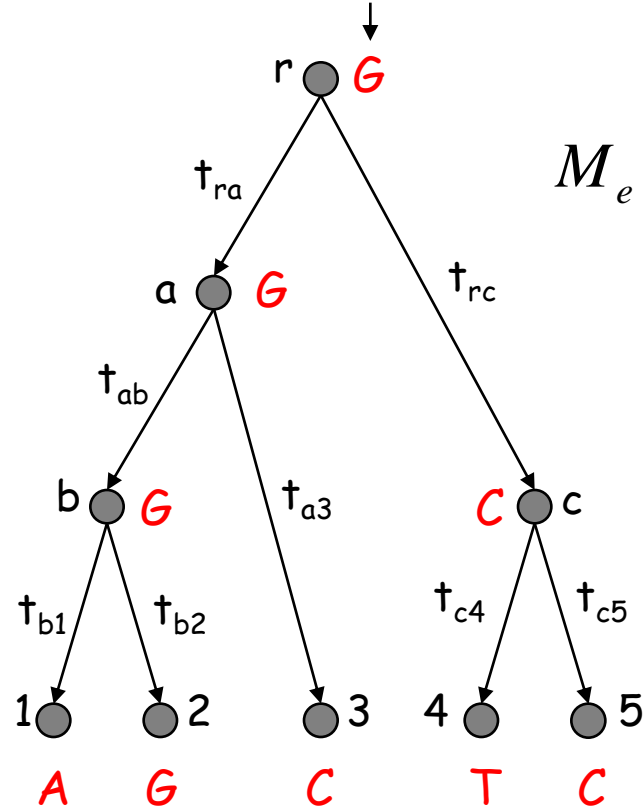
- phylogeny: T
- rate of mutation: u
- number of species: n
- number of states: $r (=4)$

$$Q = \begin{pmatrix} & A & & G \\ & \leftarrow u/3 & & \rightarrow u/3 \\ u/3 & \uparrow & & \downarrow u/3 \\ & C & & T \\ & \leftarrow u/3 & & \rightarrow u/3 \\ u/3 & \downarrow & & \uparrow u/3 \end{pmatrix}$$

- note

- no deletion/insertion
- assume molecular clock

...AACTAAGGTCGCAGGGCCAAATG...



$$M_e = e^{t_e Q}$$

phylogenetic tree reconstruction

- **setup**

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

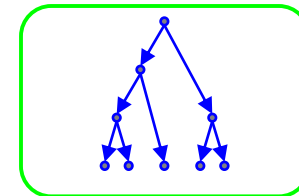
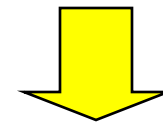
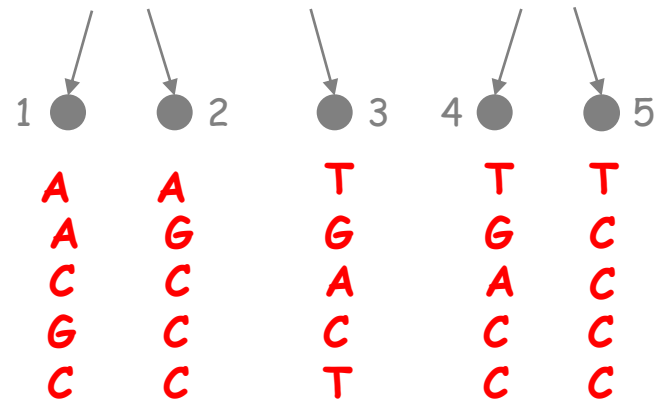
$$\Psi_n : \left\{ s_L^i \right\}_{i=1}^k \mapsto T \in T_n$$

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

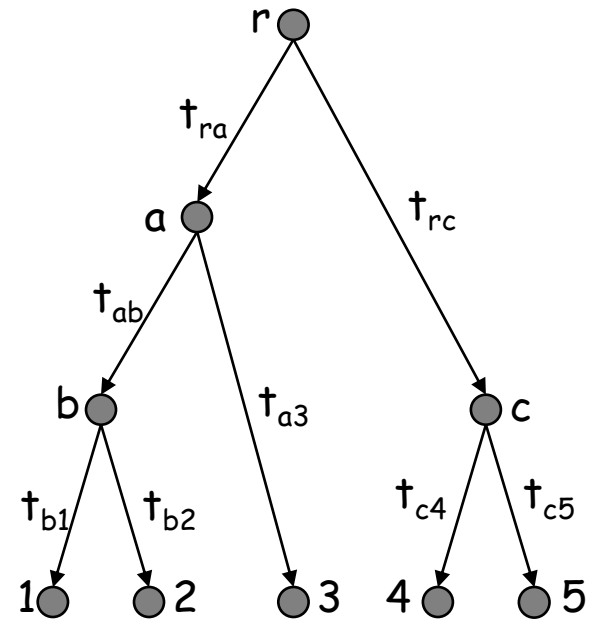


distance-matrix methods

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

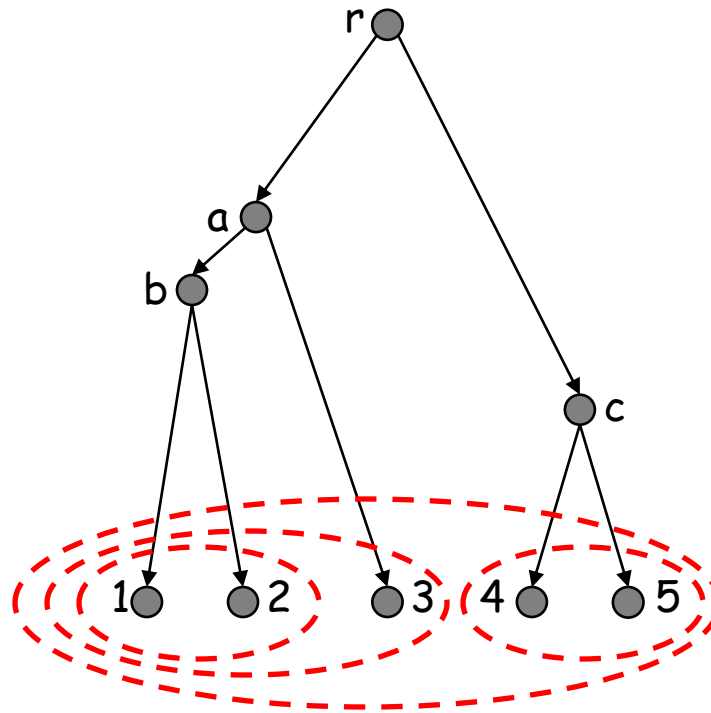
$$D(i, j) = \sum_{e \in P(T; i, j)} t_e$$

- 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



clustering algorithm

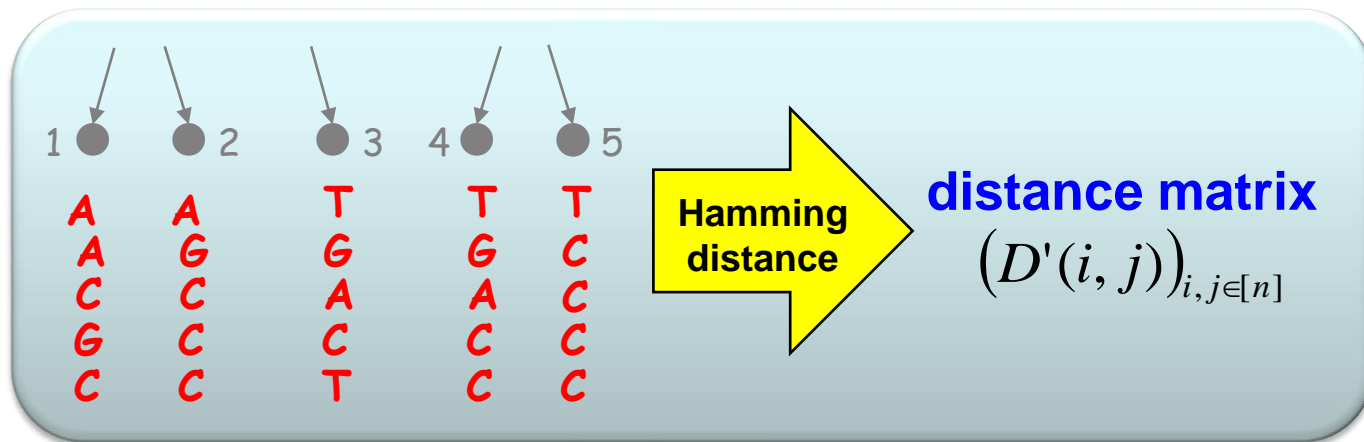
distance estimates: $(D'(i, j))_{i, j \in [n]}$



sample complexity

- **assumption (A)** - assume $0 < f < t(e) < g$, for all e
- **theorem** - UPGMA needs polynomial-length sequences, i.e.,

$$k \propto n^c$$

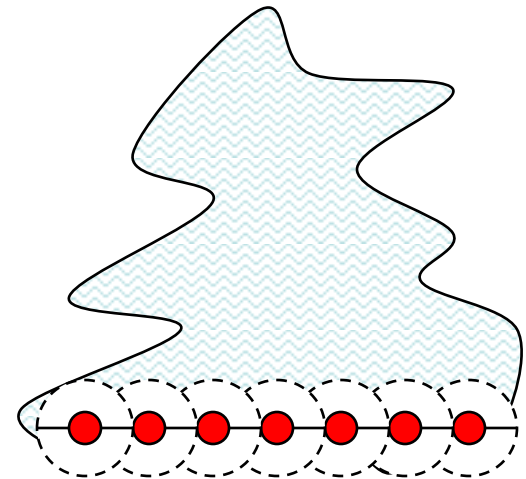


$$\frac{H(S(i), S(j))}{k} = \frac{3}{4} \left(1 - e^{-D'(i, j)}\right) \approx \frac{3}{4} - \frac{1}{n^c} \text{ if } D(i, j) = O(\log n)$$

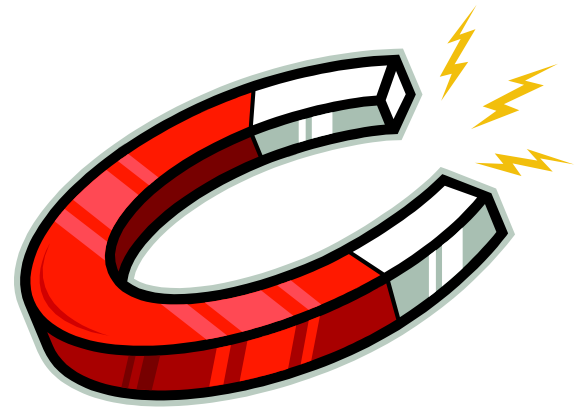
recap

- **summary**: phylogenies can be inferred in polynomial time from polynomial length sequences (general result by [Erdos-Steel-Szekely-Warnow'98])
- **question**: is this the best we can do?
- **counting argument**: need at least $\Omega(\log n)$ samples...

$$2^{O(n \log n)} \approx 2^{nk}$$

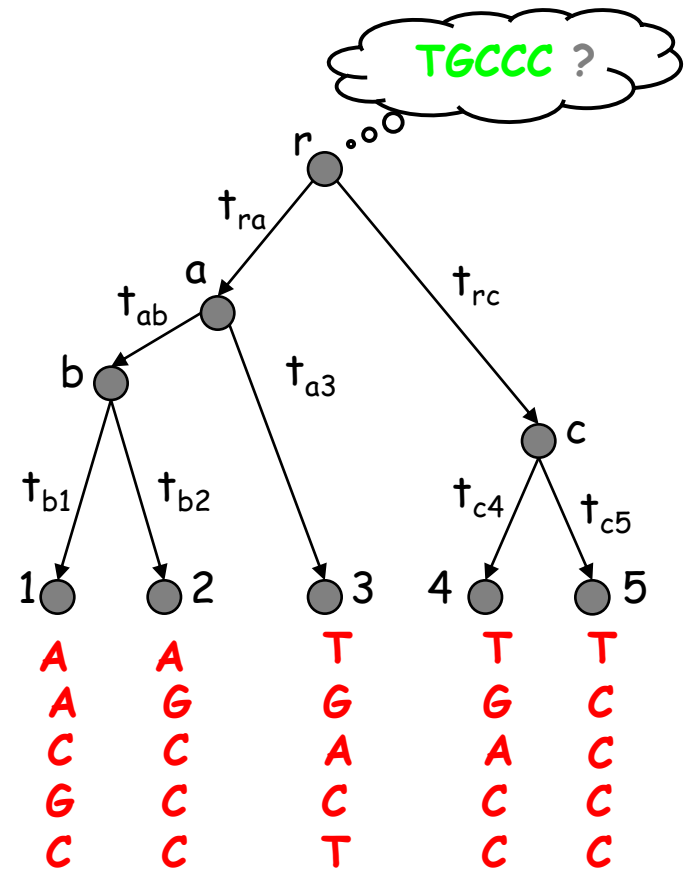


PART I
insights from
statistical physics



inferring ancestral sequences

- ancestral sequence reconstruction (a.k.a. the “reconstruction problem”)
 - **given**: sequences at leaves
 - **goal**: infer sequence at internal node
- in statistical physics terms, when is the limiting (free) Gibbs distribution **pure (or extremal)**?



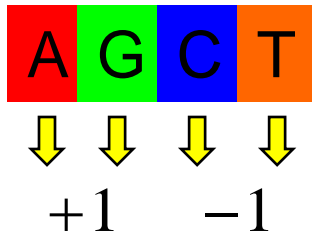
Kesten-Stigum bound

- Kesten-Stigum bound [Kesten-Stigum'66, Higuchi'77, Evans et al.'00, Mossel-Peres'03]

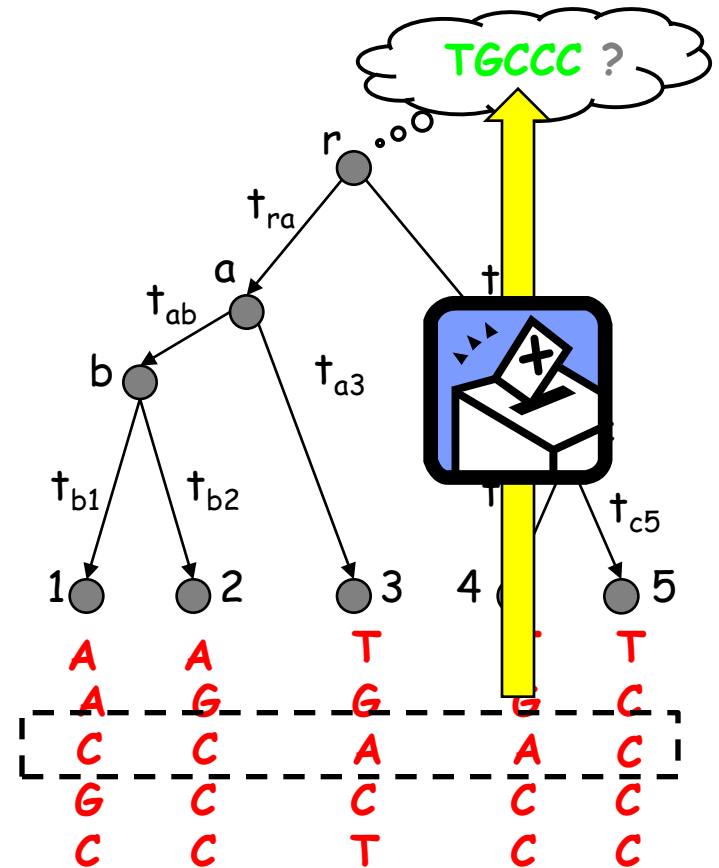
- bound on critical branch length:

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

- root estimator:

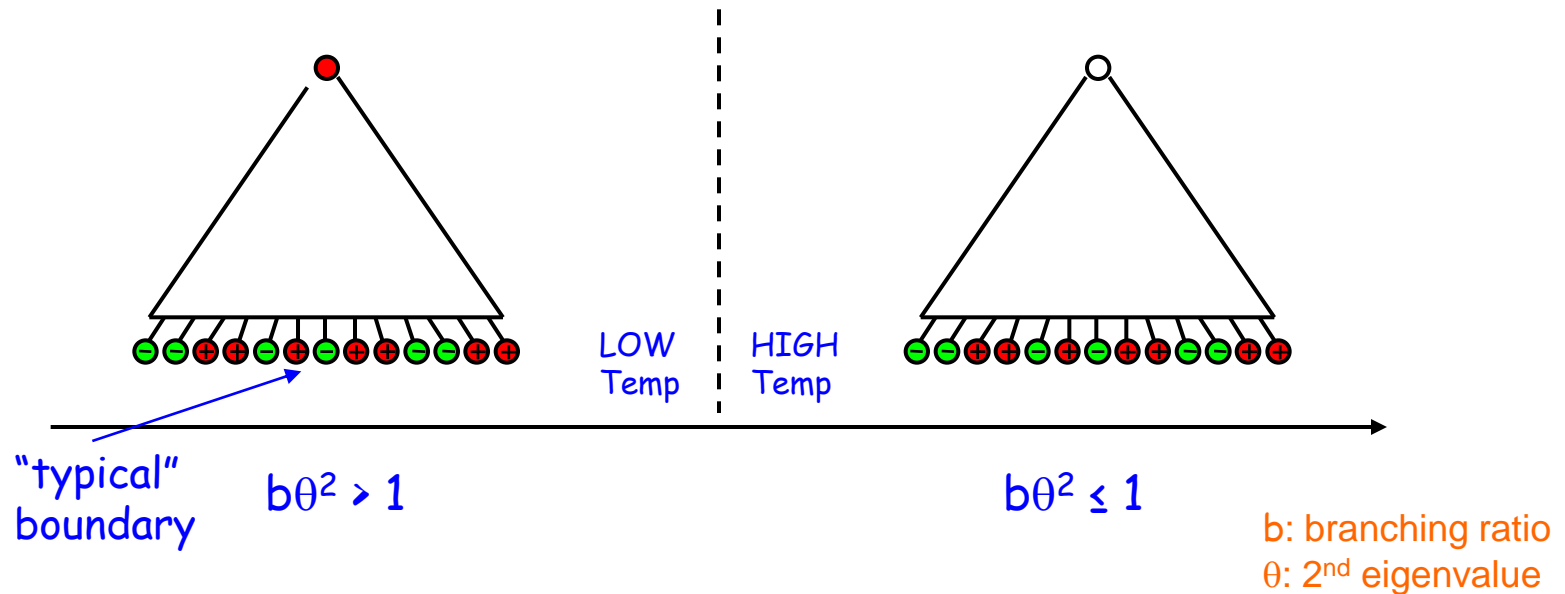


$$S = \sum_{a \in [n]} 2^{-|a|} s_a^i$$

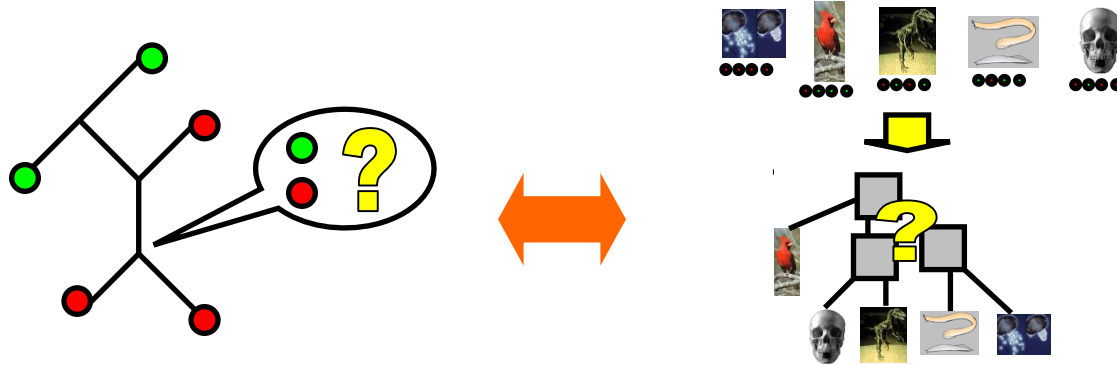


Ising case

- **theorem** - transition at $b\theta^2 = 1$
 - [Bleher-Ruiz-Zagrebnoy'95], [Ioffe'96], [Evans-Kenyon-Peres-Schulman'00], [Kenyon-Mossel-Peres'01], [Martin'03], [Martinelli-Sinclair-Weitz'04], [Borgs-Chayes-Mossel-R'06]
- reconstruction for $b\theta^2 > 1$ proved by [Higuchi'77], [Kesten-Stigum'66]
- “spinglass” case studied by [Chayes-Chayes-Sethna-Thouless'86]



Steel's conjecture: Ising model



ancestral
reconstruction

phylogenetic
reconstruction

[Daskalakis-
Mossel-R'06]

reconstruction



seq. length = $c \log n$

[Mossel'04]

non-reconstruction

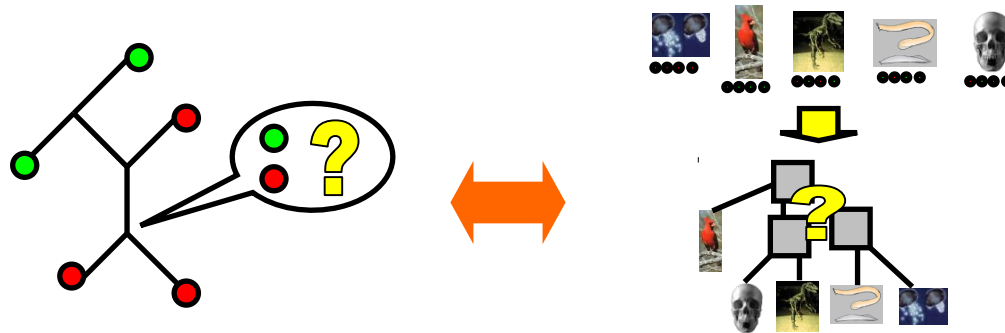


seq. length = n^c

$n = \# \text{ species}$

beyond the KS transition?

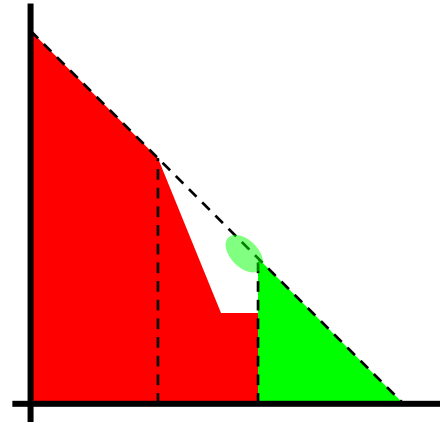
- **more recent results** [R'08, Mossel, R, Sly'08]:
 - logarithmic reconstruction possible for general 'channels' in KS zone
- **more generally, for q-state Potts model:**
 - extremality threshold need **not** correspond to KS transition -> proved by [Mossel'01] for q large enough
 - connection to **spinglasses** [Mezard, Montanari'06]: KS threshold tight iff 'dynamical' phase transition is continuous
 - results recently confirmed by [Sly'08]: KS tight for q=2,3 and not tight for $q > 4$



PART II

beyond the KS transition:

Potts model with large number of colors



Steel's conjecture: continued

- **assumption (A)** - assume $0 < f < t(e) < g$, for all e

- **theorem** [Mossel, R, Sly'08] - for all

$$\ln \sqrt{2} < g < \ln 2$$

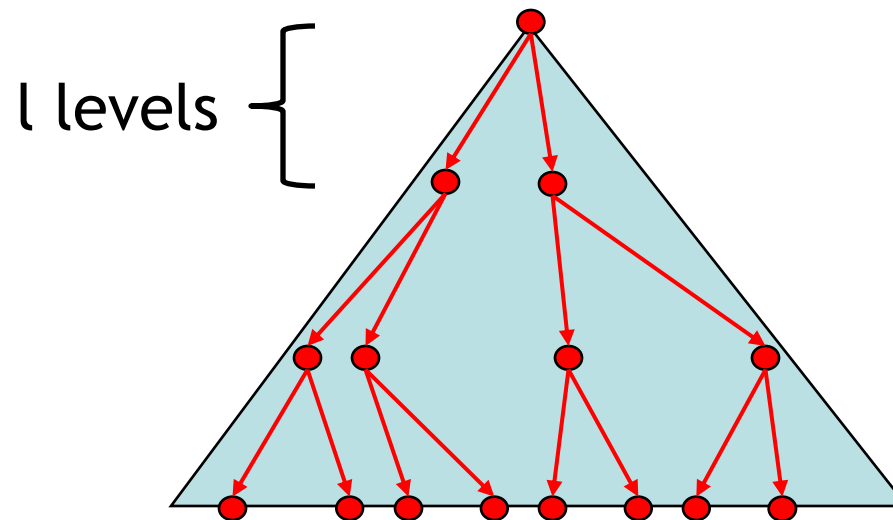
there is $q > 0$ large enough such that phylogenetic reconstruction under (A) is possible with

$$k \propto \log n$$

- **why care?**
 - protein sequence evolution
 - sample complexity of distance-based methods
- **what it entails** - finding a root estimator that **doesn't simply 'count' the states at the boundary**

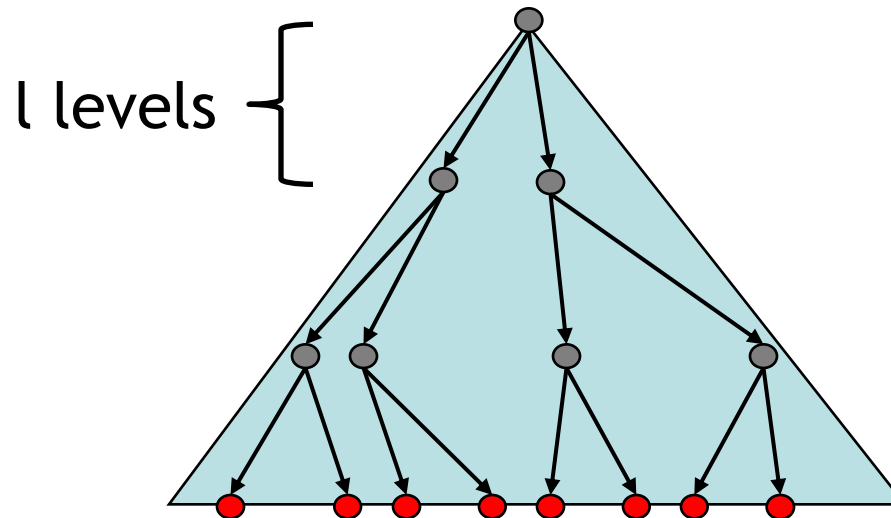
new root estimator I

- **definition** - an **l -diluted subtree** T' of T is such that, for all s , each vertex of T' on level sl of T has exactly two descendants on level $(s+1)l$



new root estimator II

- **definition** - our new **root estimator** is as follows:
 - $B(j,l)$ is the event that there is l -diluted tree with all leaves colored j
 - $B[l]$ is the set of colors j such that $B(j,l)$ holds
 - pick J u.a.r. in $1, \dots, q$
 - estimator $X = J$ if J in $B[l]$, o.w. pick uniformly in remaining colors



new root estimator III

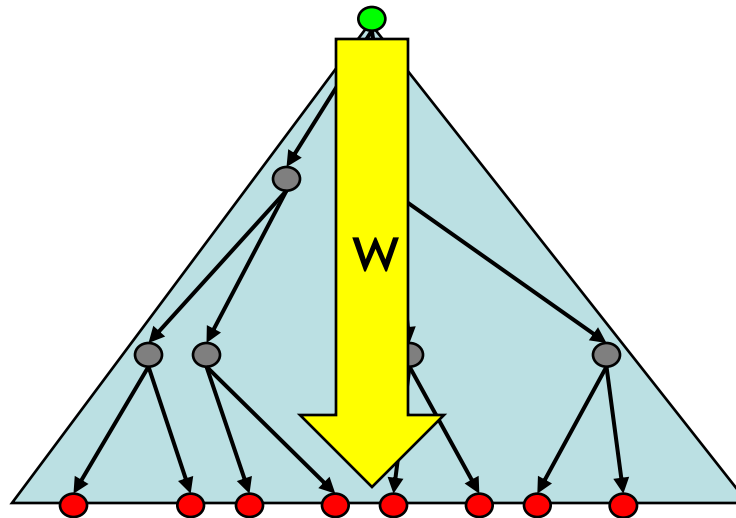
- **property** - for all

$$\ln \sqrt{2} < g < \ln 2$$

there is $q, l, W > 0$ such that the root estimator X satisfies

$$\mathbb{P}[X = j \mid \text{root} = i] = \left(e^{wQ}\right)_{ij}$$

for all colors i, j , where moreover $w < W$



future directions

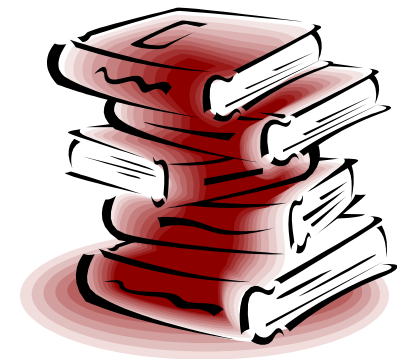
- future directions
 - sample complexity of MLE and parsimony



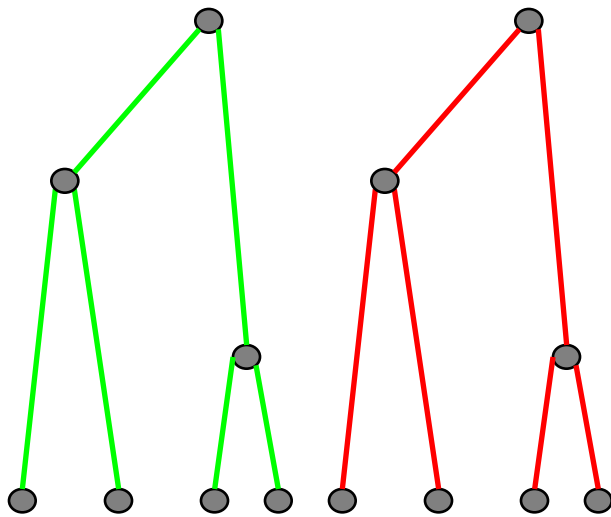
thank
you

references

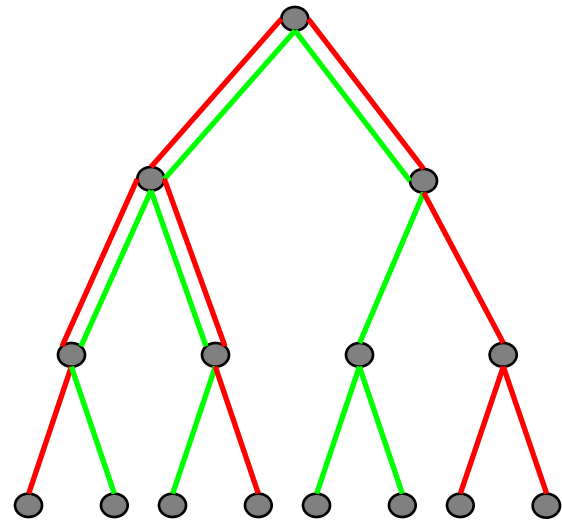
- R, “A Short Proof that Phylogenetic Tree Reconstruction by Maximum Likelihood is Hard”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2006
- Mossel, R, “Learning Nonsingular Phylogenies and Hidden Markov Models”, *Annals of Applied Probability*, 2006
- Daskalakis, Mossel, R, “Optimal Phylogenetic Reconstruction”, *Proceedings of ACM STOC’06*
- Borgs, Chayes, Mossel, R, The Kesten-Stigum Bound is Tight for Roughly Symmetric Channels, *Proceedings of IEEE FOCS’06*
- Daskalakis, Mossel, R, “Phylogenies Without Branch Bounds: Contracting the Deep, Pruning the Deep”, *Preprint*, 2007



edge disjointness

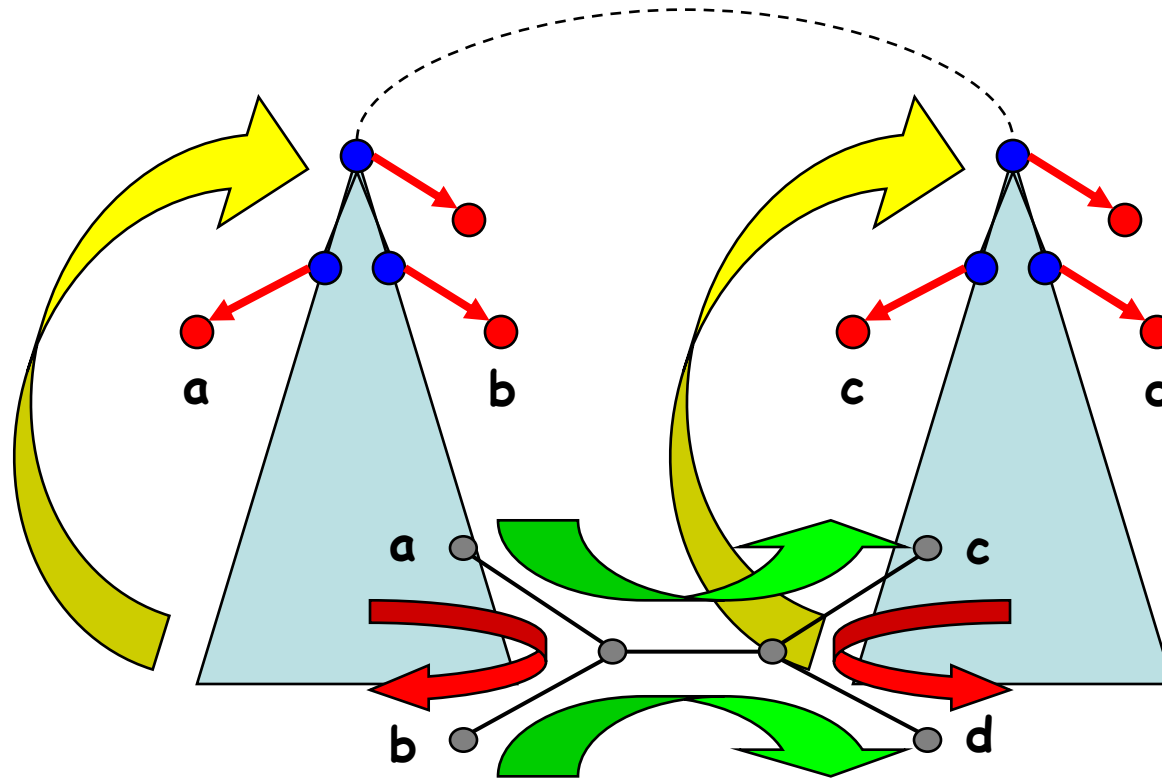


non edge-disjoint reconstruction

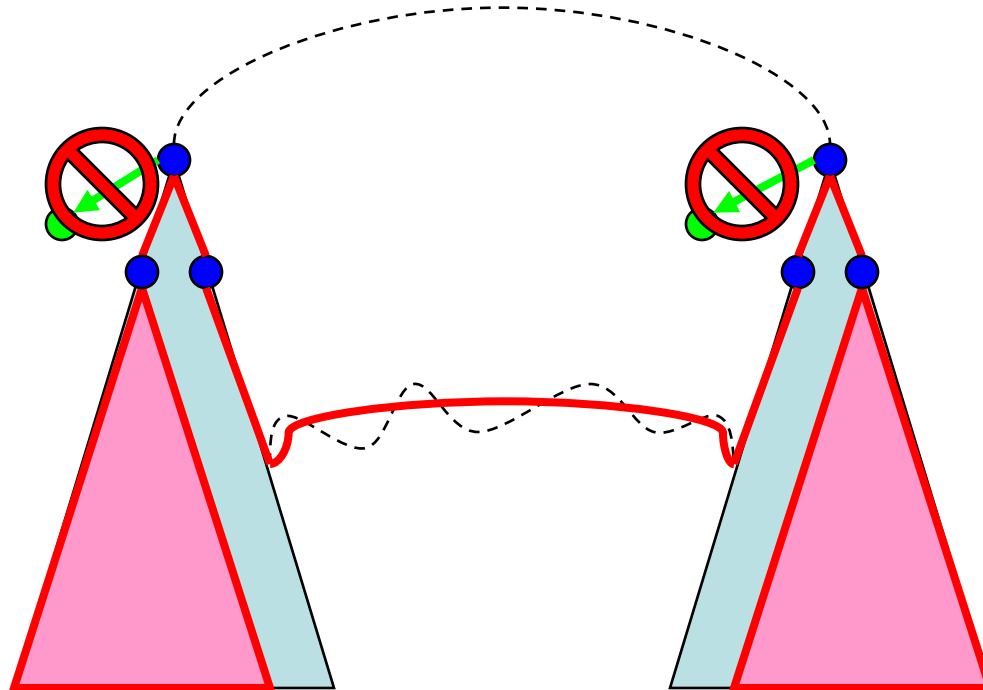


true tree

estimating internal distances I

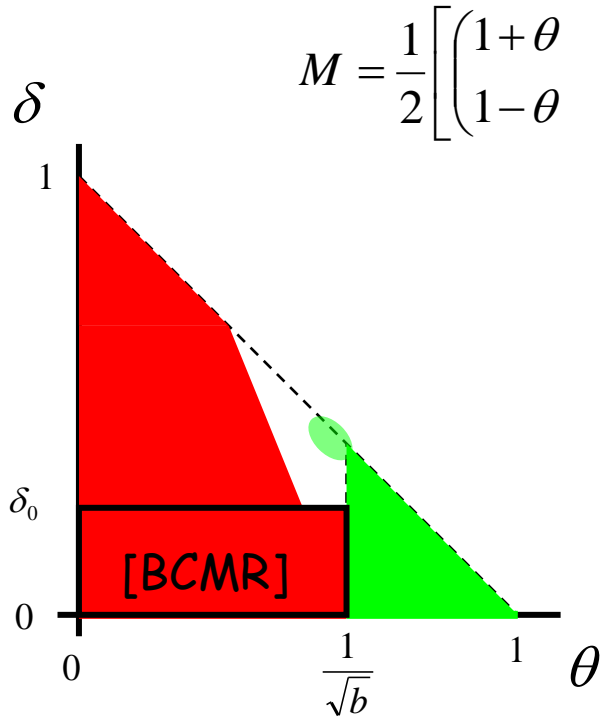


estimating internal distances II

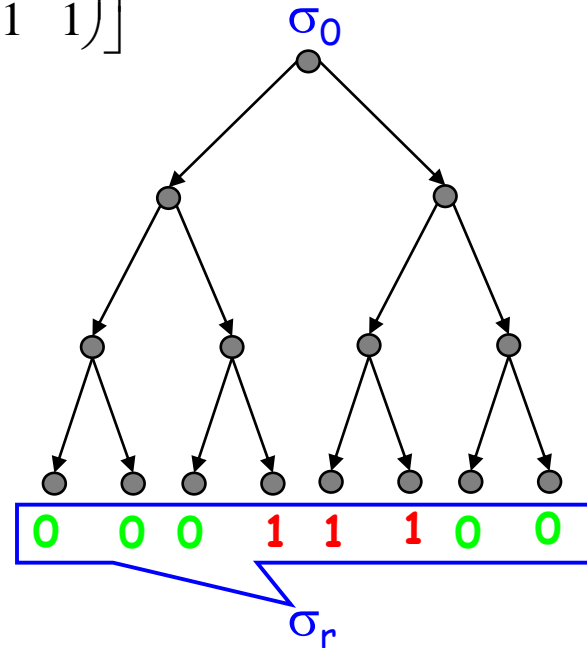


recall

- theorem** [Borgs-Chayes-Mossel-R'06] - exists $\delta_0 > 0$ s.t. if $b\theta^2 \leq 1$ and $|\delta| < \delta_0$ then the reconstruction problem is **not solvable**



$$M = \frac{1}{2} \left[\begin{pmatrix} 1+\theta & 1-\theta \\ 1-\theta & 1+\theta \end{pmatrix} + \delta \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} \right]$$



magnetization of the root

- we use $\{+1, -1\}$.
- stationary distribution of channel M

$$\pi = (\pi_+, \pi_-)$$

- **definition** - the **magnetization of the root** is

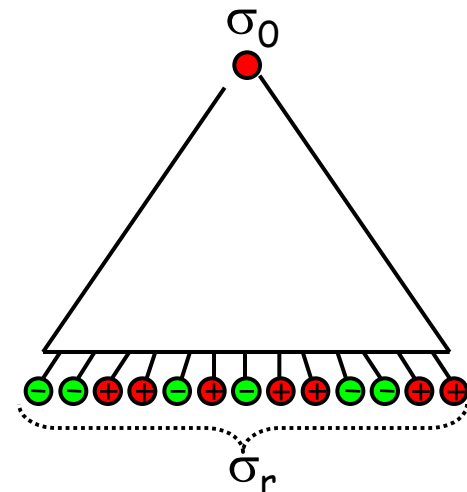
$$X_r(\sigma_r) = \pi_-^{-1} \{ \pi_- P[\sigma_0 = +1 | \sigma_r] - \pi_+ P[\sigma_0 = -1 | \sigma_r] \}$$

- **lemma** - it suffices to show

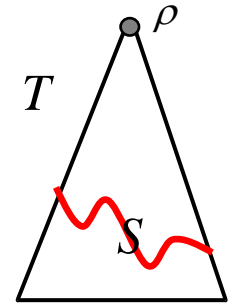
$$\bar{x}_r \equiv E_{T_r}^\pi [X_r^2] \rightarrow 0$$

- basic proof idea: **moment recursion**

$$\bar{x}_r \leq b\theta^2 \bar{x}_{r-1}$$



more general result



- general trees - previous results

- **definition** - the **branching number** is defined as

$$\text{br}(T, \theta) = \inf \left\{ \lambda > 0 : \inf_{\text{cutsets } S} \sum_{x \in S} \left(\lambda^{-|x|} \prod_{e \in \text{path}(\rho, x)} \theta^2(e) \right) = 0 \right\}$$

- [Evans-Kenyon-Peres-Schulman'00] binary symmetric case on general tree, solvable iff $\text{br}(T, \theta) > 1$

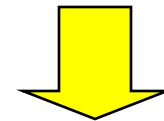
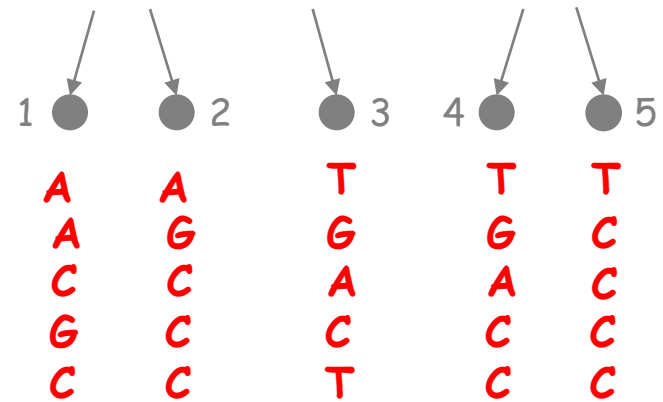
- **theorem** [Borgs-Chayes-Mossel-R'06] - let $0 \leq \theta_0 < 1$. exists $\delta_0 > 0$ such that

- for all stationary distributions $\pi = (\pi_+, \pi_-)$ with $\max\{|\delta(\pi, \theta)|, |\delta(\pi, -\theta)|\} < \delta_0$
- for all trees with $\sup_e |\theta(e)| \leq \theta_0$ and $\text{br}(T, \theta) \leq 1$

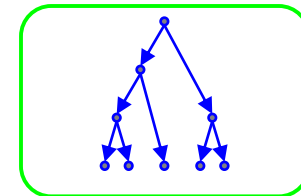
the reconstruction problem is **not solvable**

theoretical approach

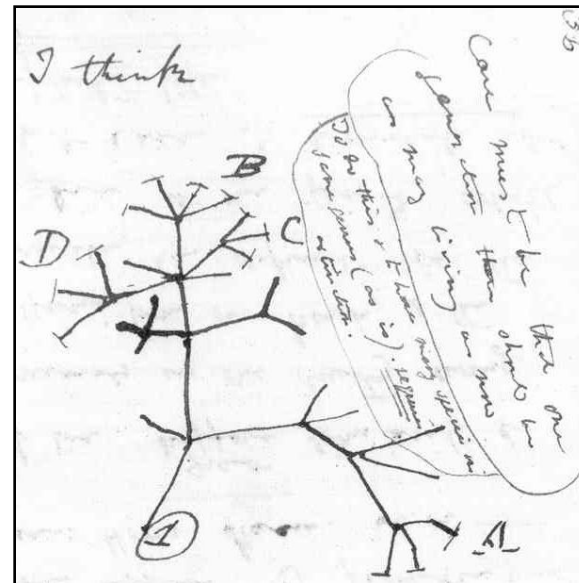
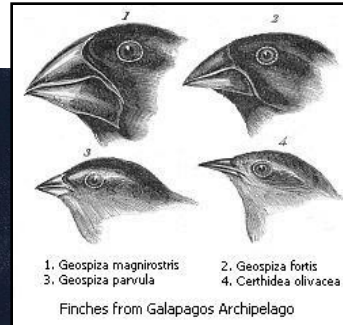
- two different approaches:
 - arbitrary dataset
 - model-generated dataset
- we focus on the latter
- plus: want “efficient” reconstruction



$n = \# \text{ species}$
 $k = \text{seq. length}$



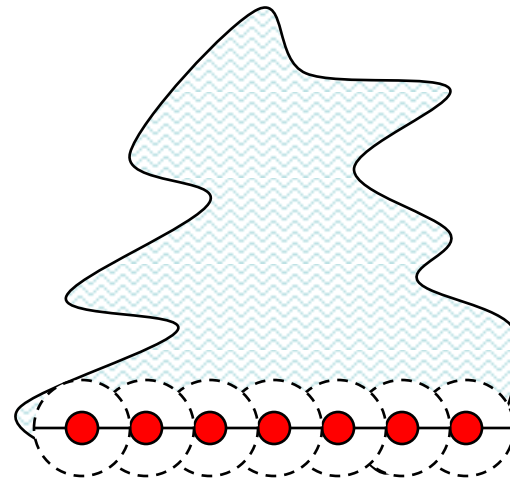
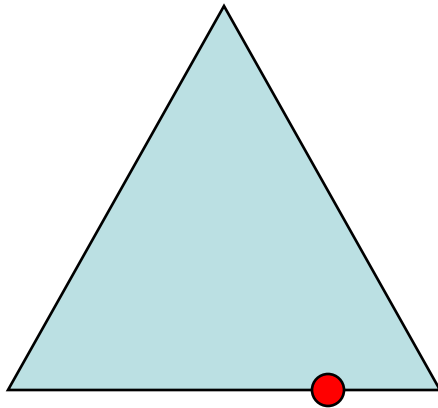
Darwin's finches



“very local” metric

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

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



maximum likelihood

- **data:** n $\{0,1\}$ -sequences of length k

$$\{S(j) = (s_L^1(j), \dots, s_L^k(j)) \in \{0,1\}^k : 1 \leq j \leq n\}$$

- **likelihood**

$$\Lambda(T, \{t_e\}; S(1), \dots, S(n)) = \prod_{i=1}^k \sum_{s^* \in \text{Ext}(s_i)} \prod_{e=(u,v) \in E} p(t_e)^{\langle s^*(u) \neq s^*(v) \rangle} (1 - p(t_e))^{\langle s^*(u) = s^*(v) \rangle}$$

- **MLE**

$$(T^*, \{t_e^*\}) = \arg \min_{(T, \{t_e\})} [-\ln \Lambda(T, \{t_e\}; S(1), \dots, S(n))]$$

- statistically consistent [Chang'96], but:
 - **theorem** [Chor-Tuller'06, R'06] - NP-hard (i.e. “computationally intractable”); actually hard to approximate