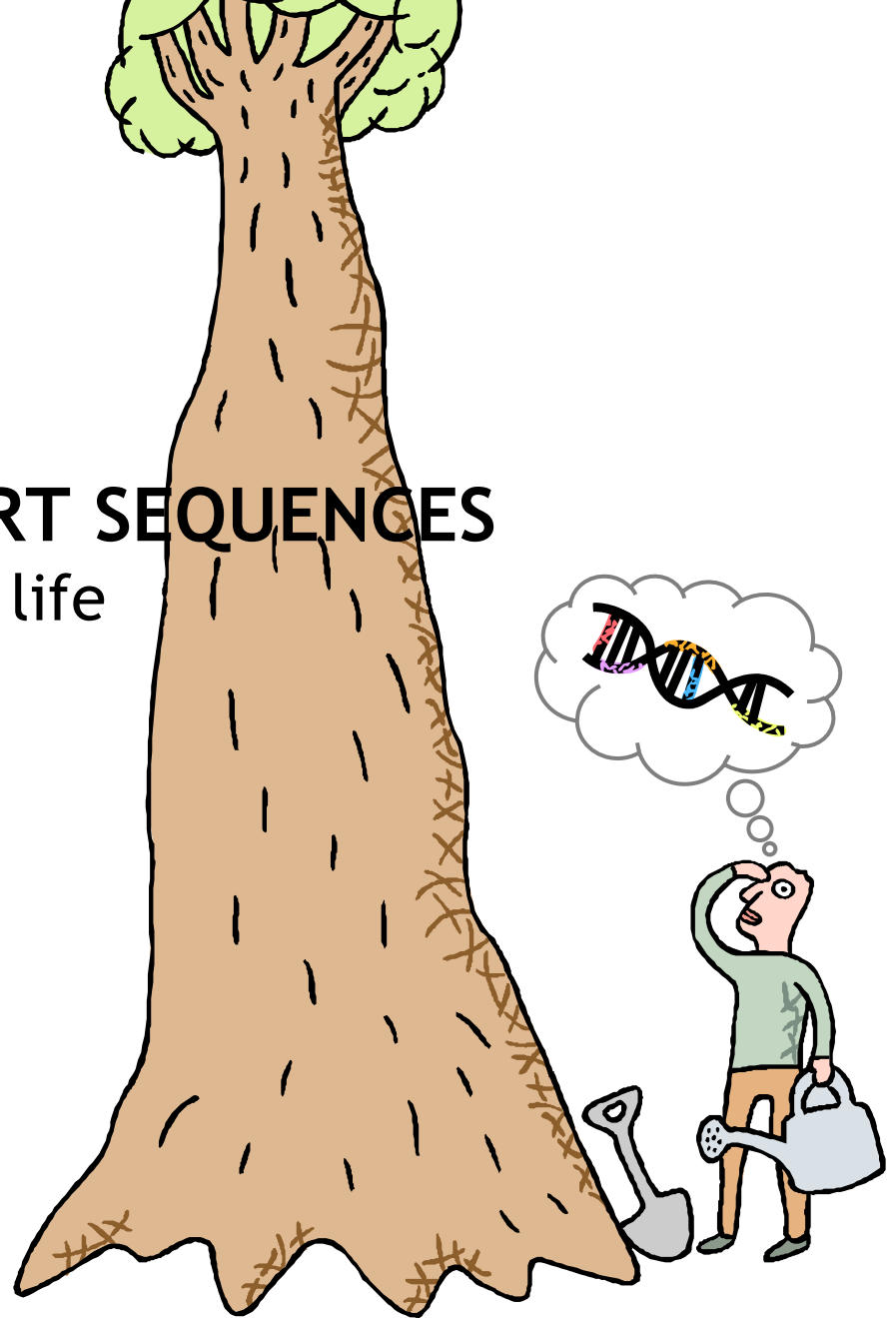


BIG TREES FROM SHORT SEQUENCES

reconstructing the tree of life

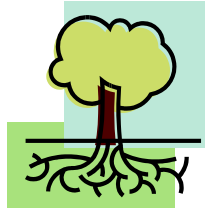
Sebastien Roch
Microsoft Research

collaborators:
C. Borgs, J. Chayes, C. Daskalakis,
E. Mossel, M. Steel

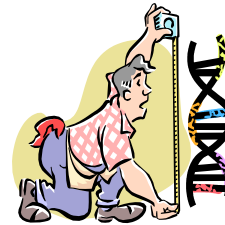


outline of the talk

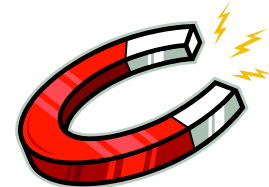
PART 0
background



PART I
revisiting
distance matrix
methods

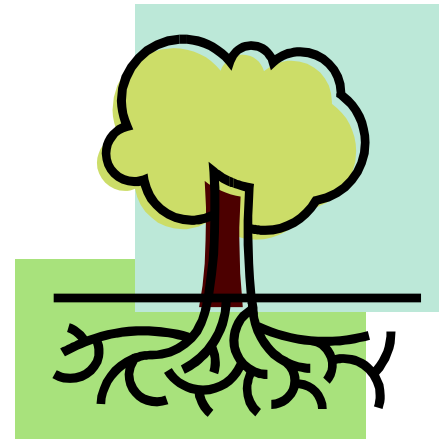


PART II
insights from
statistical physics



PART 0-a

background: Markov models on trees



Markov chain on a tree

- **broadcasting model**

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

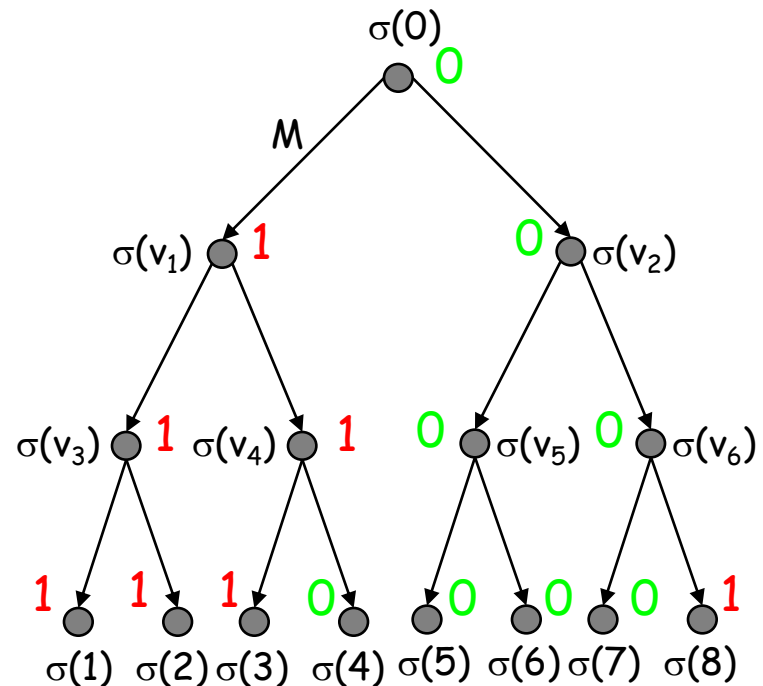
$$\{\sigma(v) \in \{0,1\} : v \in V\}$$

- Markov transition matrix:

$$M = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

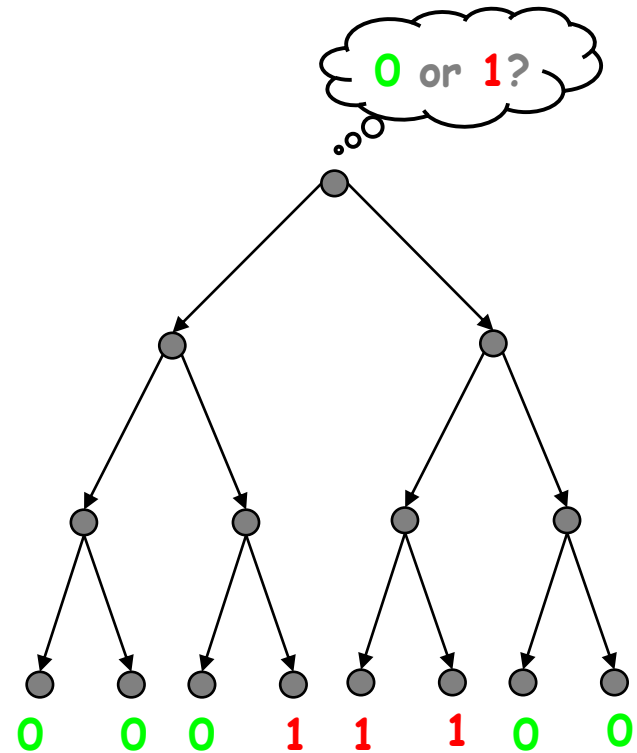
- stationary distribution:

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



the reconstruction problem

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



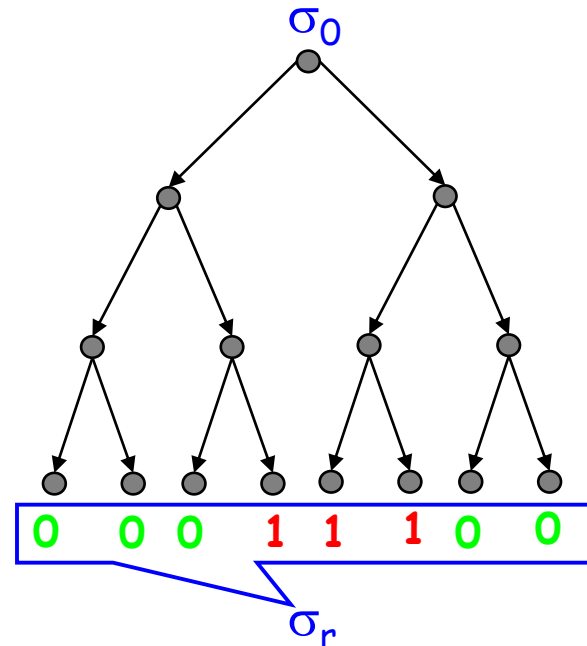
reconstruction solvability

- **setup**

- T : infinite rooted b-ary tree
- T_r : first r levels of T
- M : Markov transition matrix

- **definition** - the reconstruction problem on (T, M) is **solvable** if the following condition holds:

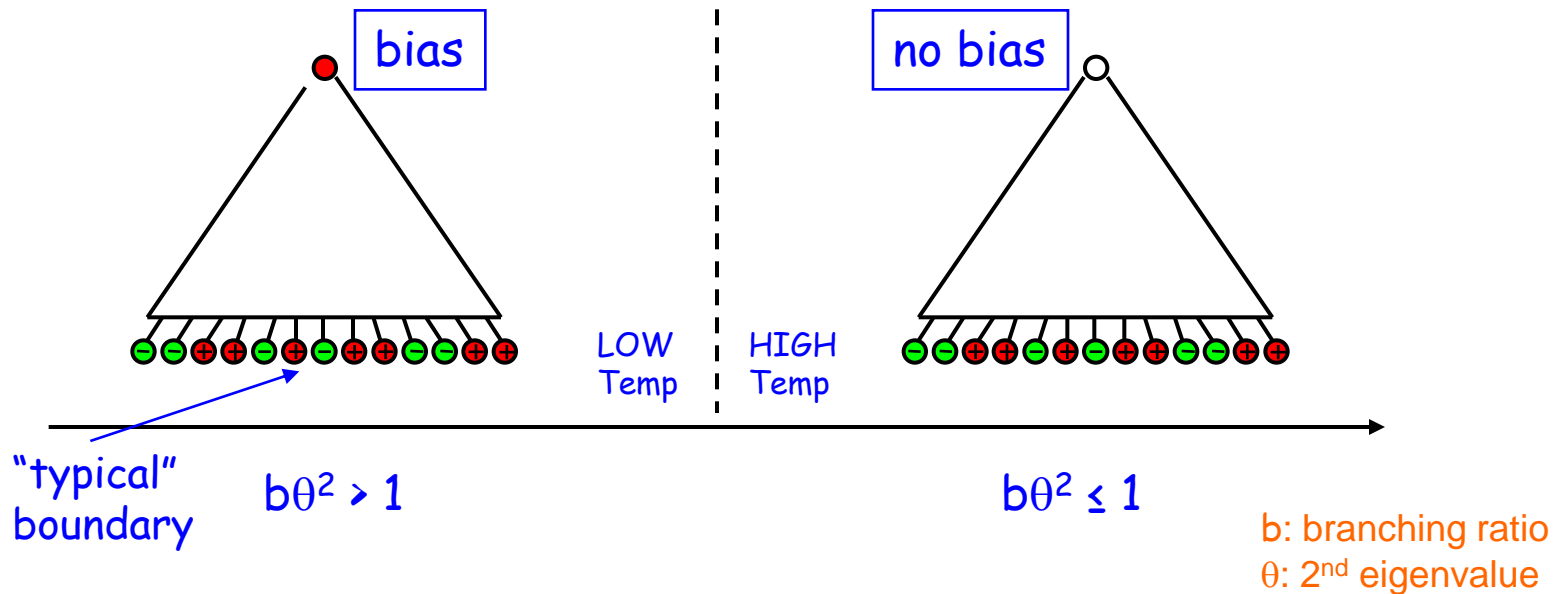
- $\lim_r |P_r^0 - P_r^1|_{TV} > 0$, where P_r^j denotes the distribution of σ_r conditional on $\sigma_0 = j$



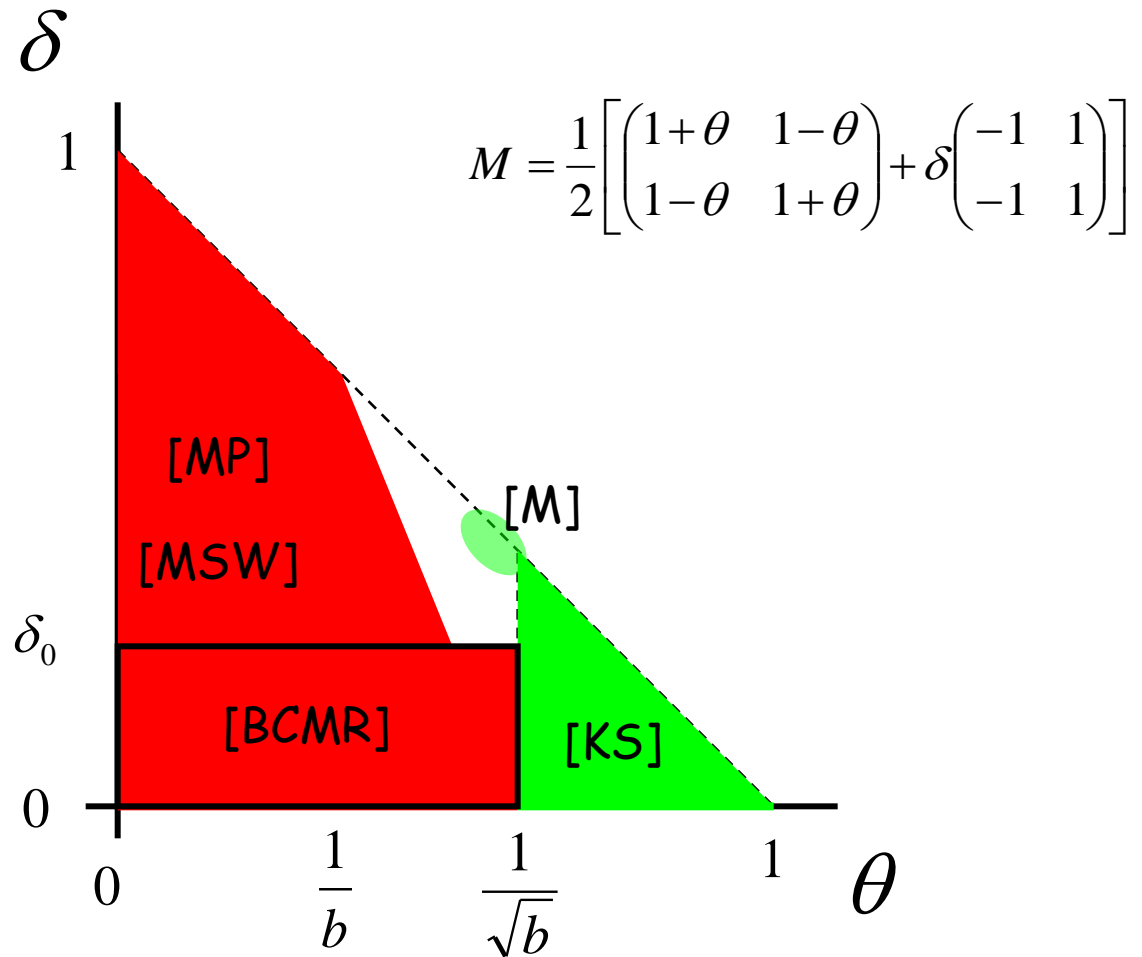
$$M = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \frac{1}{2} \left[\begin{pmatrix} 1+\theta & 1-\theta \\ 1-\theta & 1+\theta \end{pmatrix} + \delta \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} \right], \text{ where } \theta = \lambda_2(M)$$

symmetric case ($\delta=0$)

- **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]
- solvability for $b\theta^2 > 1$ proved by [Higuchi'77], [Kesten-Stigum'66]
- “spinglass” case studied by [Chayes-Chayes-Sethna-Thouless'86]

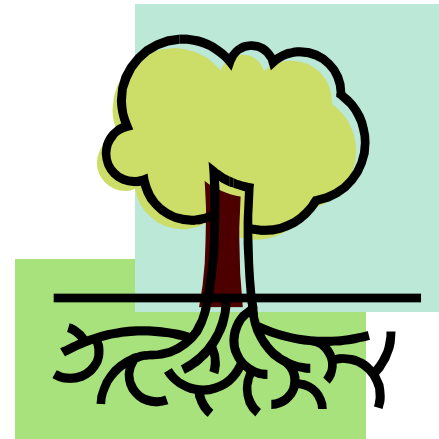


phase diagram

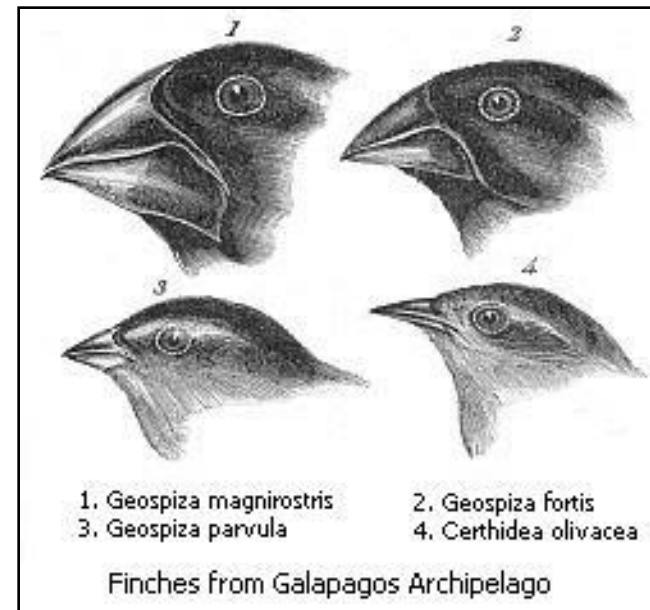
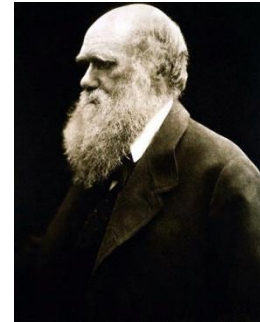


PART 0-b

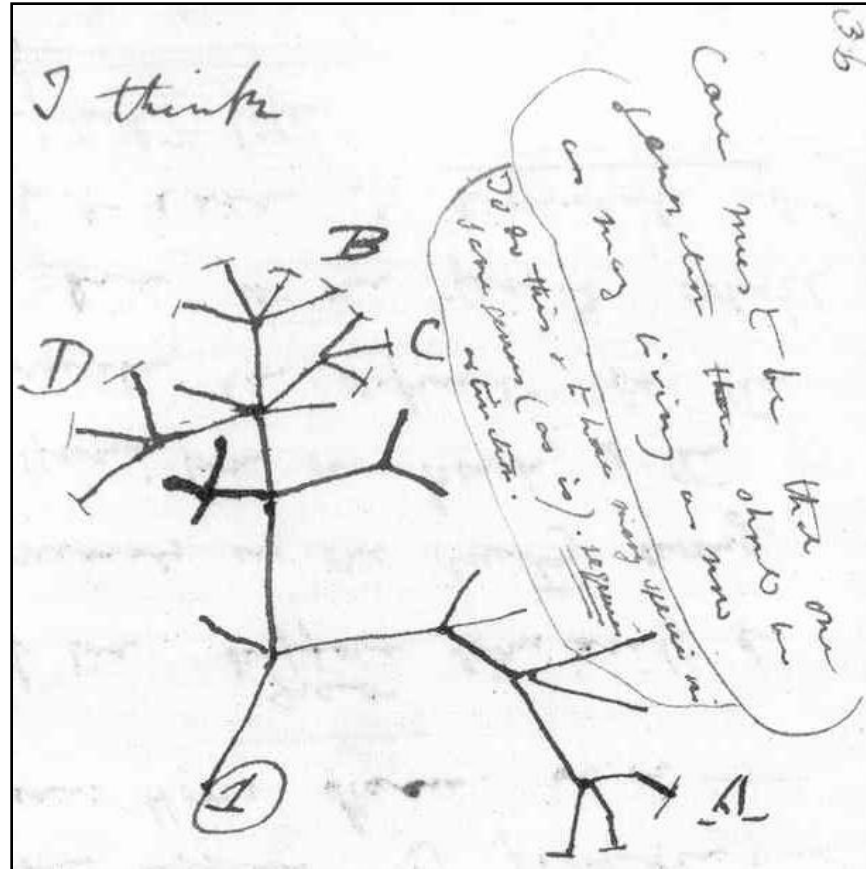
background: statistical phylogenetics



Darwin's finches



“i think”



From: Darwin, Transmutation Notebook B

DNA sequence evolution

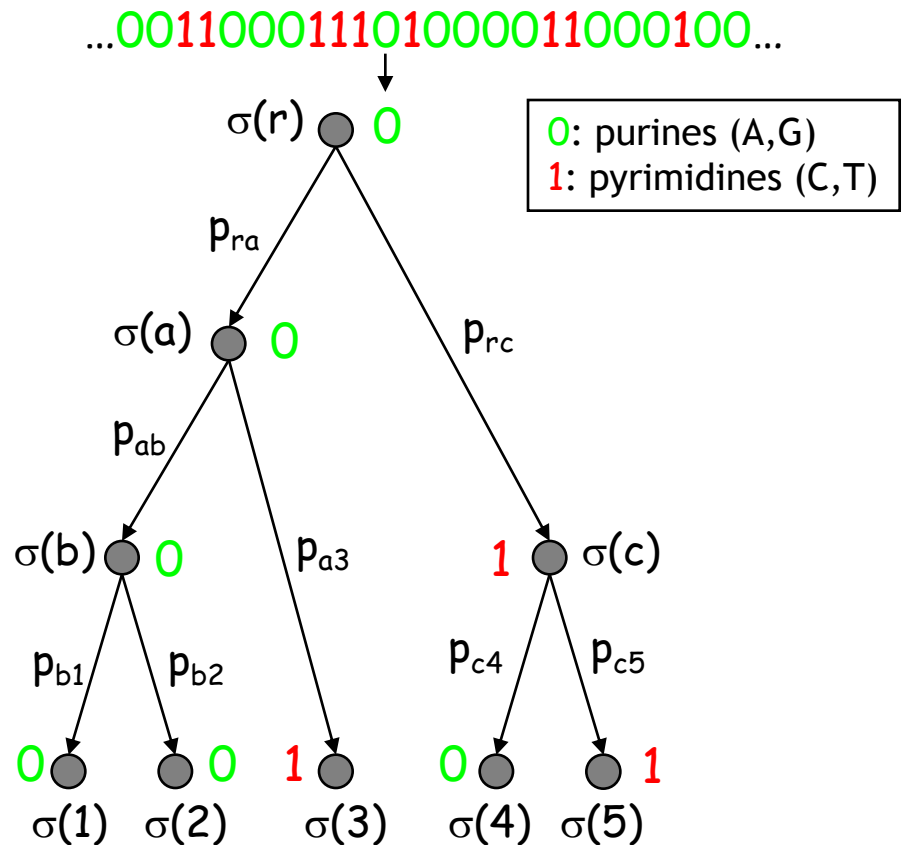
- CFN model

- only mutations
- tree: $T = (V, E)$
- node states:

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

- mutation probabilities:

$$\{0 < p_e < 1/2 : e \in E\}$$



phylogenetic tree reconstruction

- **setup**

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

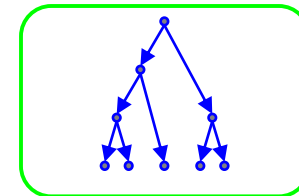
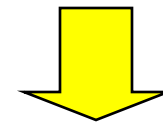
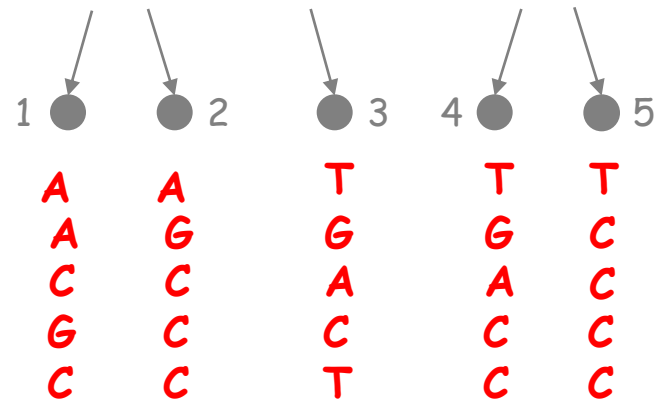
$$\Psi_n : \left\{ \sigma_L^i \right\}_{i=1}^k \mapsto T \in \mathcal{T}_n$$

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

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

- **efficiency**

- computational: running time
- information-theoretic: k



maximum likelihood

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

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

- **likelihood**

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

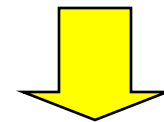
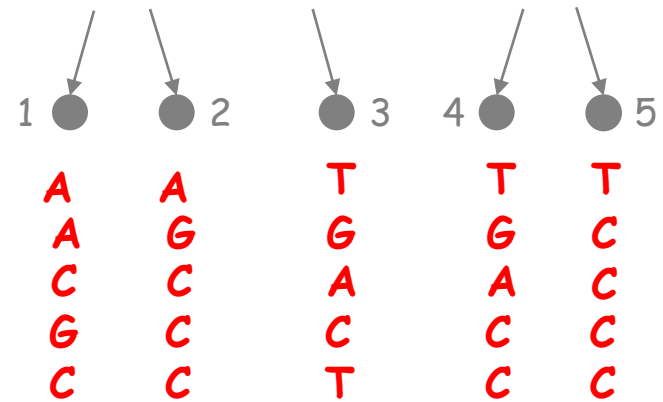
- **MLE**

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

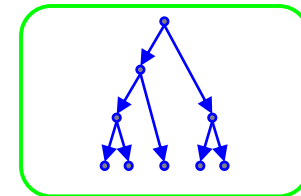
- statistically consistent [Chang'96], but:
 - **theorem** [Chor-Tuller'06, R'06] - NP-hard (i.e. “computationally intractable”); actually hard to approximate

theoretical approach

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

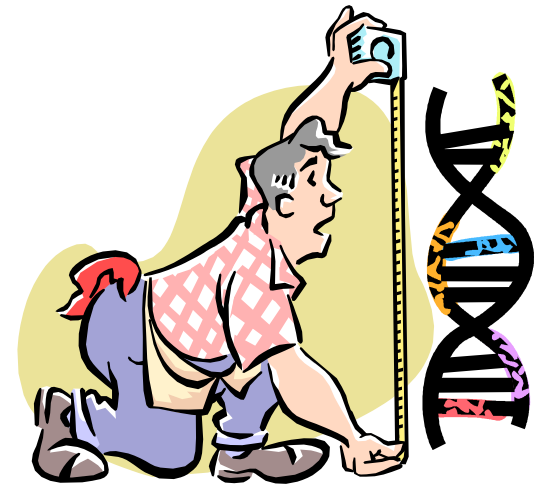


$n = \# \text{ species}$
 $k = \text{seq. length}$



PART I

revisiting distance matrix methods



distance matrix methods

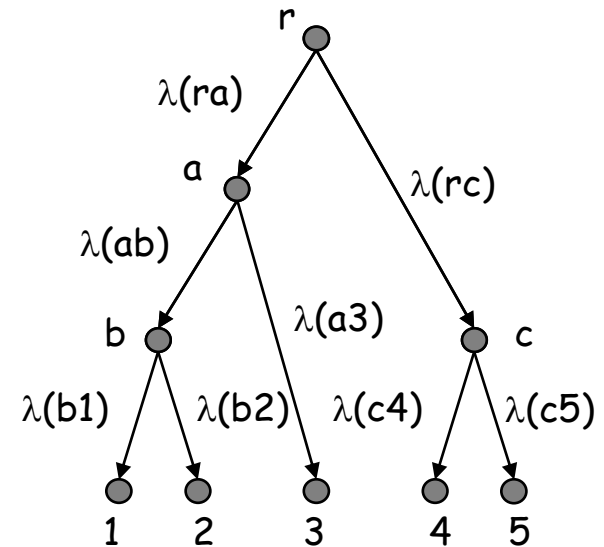
- in CFN case:
 - associate to each edge e a **weight**

$$\lambda(e) = -\ln(1 - 2p_e)$$

- defines a **tree metric**

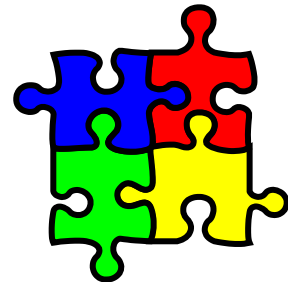
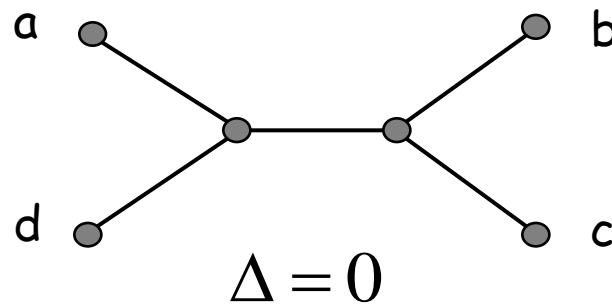
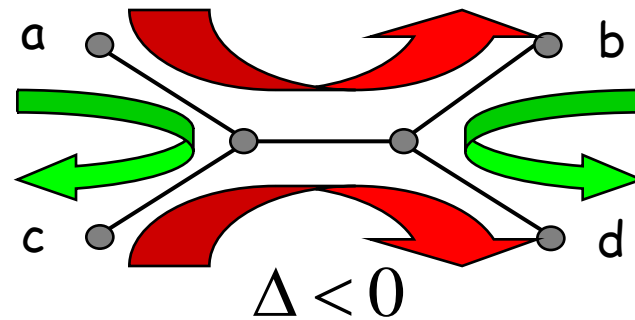
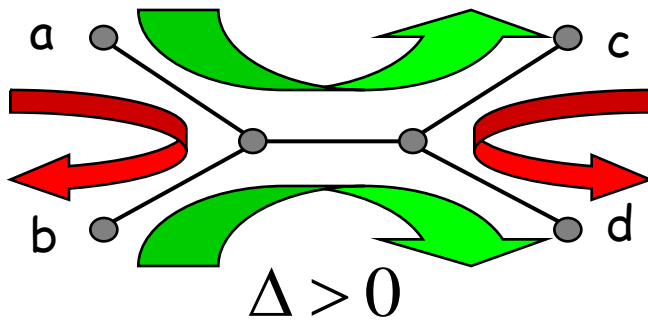
$$D(i, j) = \sum_{e \in P(T; i, j)} \lambda(e) = -\ln(1 - 2P[\sigma(i) \neq \sigma(j)])$$

- generalized by [Steel'94]
- reconstruction algorithm:
 - estimate $D(i, j)$ from sequences
 - deduce the topology of the tree
- **theorem** - reconstruction can be done efficiently (polynomial time)
 - e.g. Neighbor-Joining



four-point method

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



distance matrix methods

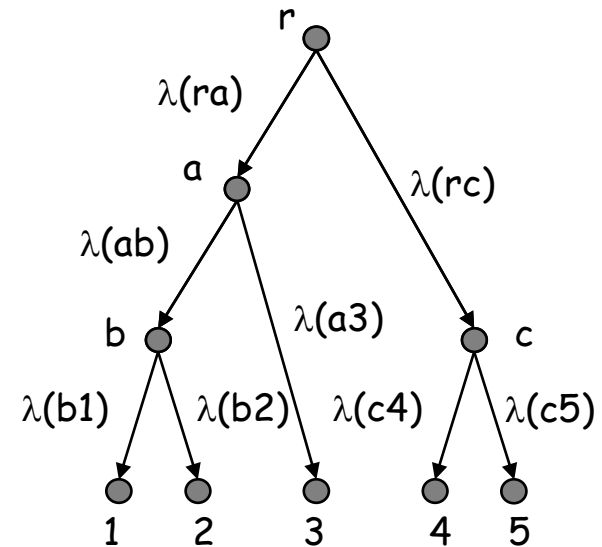
- in CFN case:
 - associate to each edge e a **weight**

$$\lambda(e) = -\ln(1 - 2p_e)$$

- defines a **tree metric**

$$D(i, j) = \sum_{e \in P(T; i, j)} \lambda(e) = -\ln(1 - 2P[\sigma(i) \neq \sigma(j)])$$

- generalized by [Steel'94]
- reconstruction algorithm:
 - estimate $D(i, j)$ from sequences
 - deduce the topology of the tree
- **theorem** - reconstruction can be done efficiently (polynomial time)
 - e.g. Neighbor-Joining



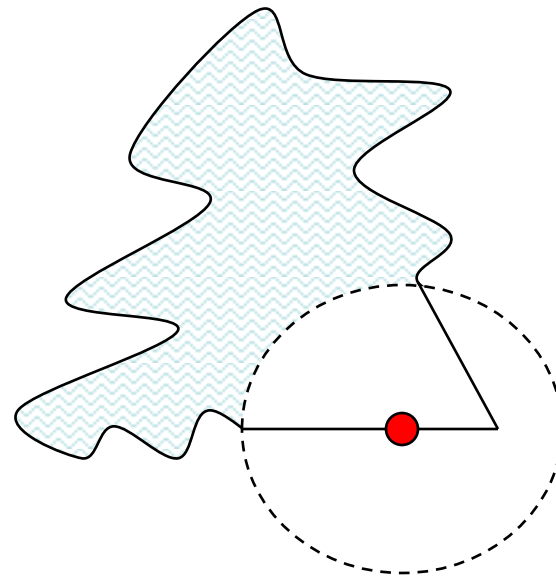
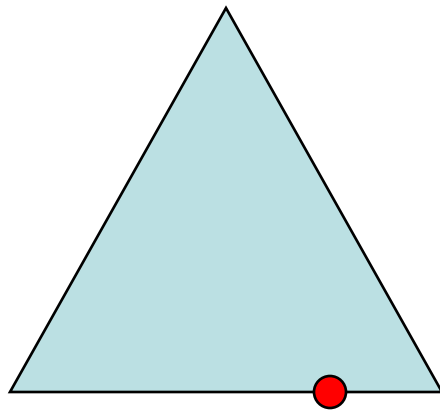
“local” metric

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

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

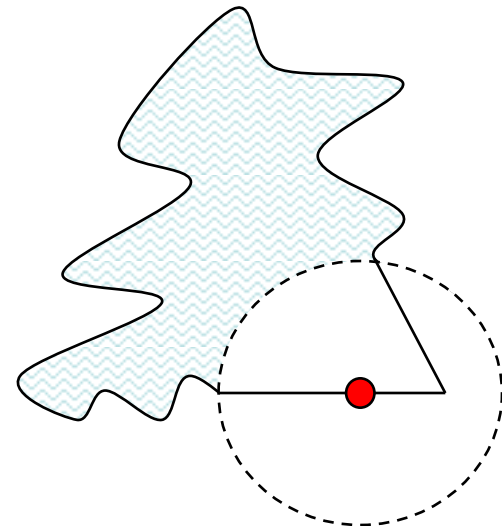
- **definition** [Mossel'07] - a symmetric matrix d is a (ε, M) -**distortion** of the distance matrix D if

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



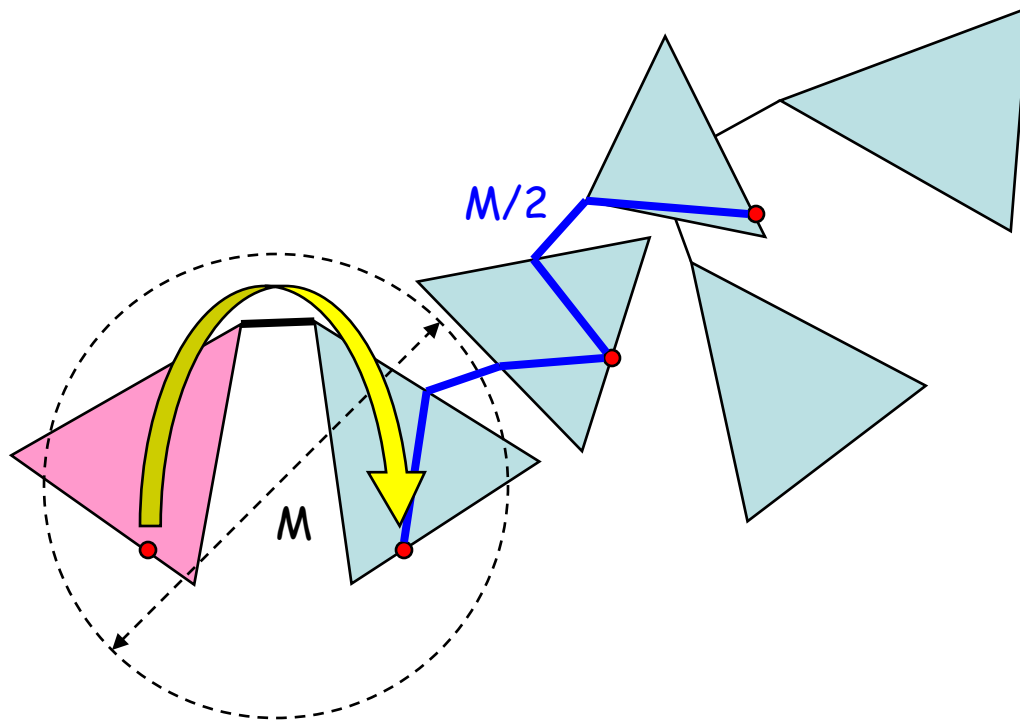
distance methods revisited

- **assumption (A)** - for simplicity assume $0 < f < \lambda(e) < g$, for all e
- **theorem** [Erdos-Steel-Szekely-Warnow'97,'98], [Mossel'07], [Daskalakis et al.'06], [Gronau-Moran-Snir'08], [Daskalakis-Mossel-R'08] - can achieve polynomial-length sequences and polynomial time



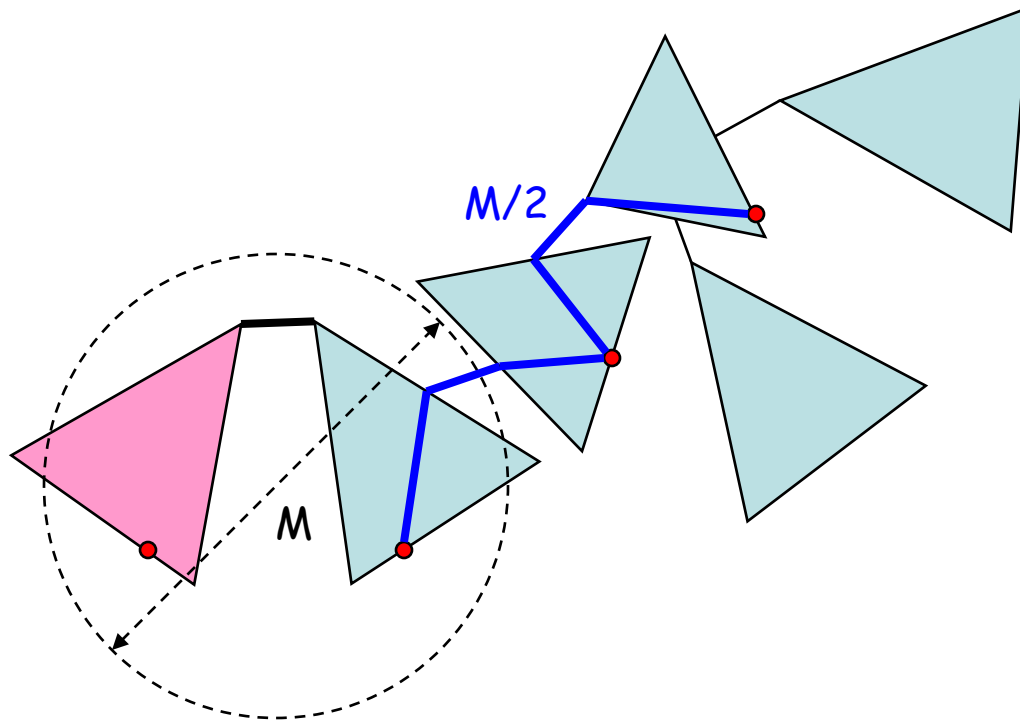
algorithm from [Daskalakis-Mossel-R'08]

- **assumption (A)** - for simplicity assume $0 < f < \lambda(e) < g$, for all e
- **observation** - with seq. length = n^γ , we can take $M = C \log n$
- **fact** - every edge is on a path of length $C' \log n$ (take $C = 10 C'$)



reconstructing contracted forests

- **assumption (A)** - for simplicity assume $0 < f < \lambda(e) < g$, for all e
- **theorem** [Daskalakis-Mossel-R'08] - without assumption (A), we recover a contracted forest



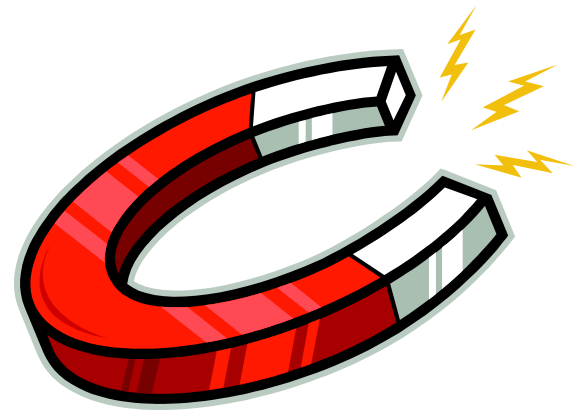
recap

- **summary**: phylogenies can be inferred in polynomial time from polynomial length sequences (transition matrices can also be estimated [Mossel-R'06])
- **question**: is this the best we can do?
- **counting argument**: need at least $\Omega(\log n)$ samples...

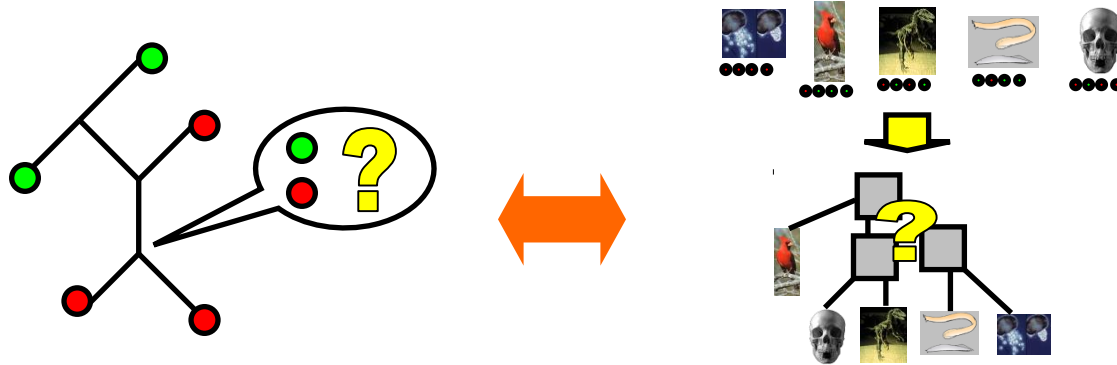
$$2^{O(n \log n)} \approx 2^{nk}$$

PART II

insights from statistical physics



resolution of Steel's conjecture



ancestral
reconstruction

phylogenetic
reconstruction

[Daskalakis-
Mossel-R'06]

short branches



seq. length = $c \log n$

[Mossel'04]

long branches



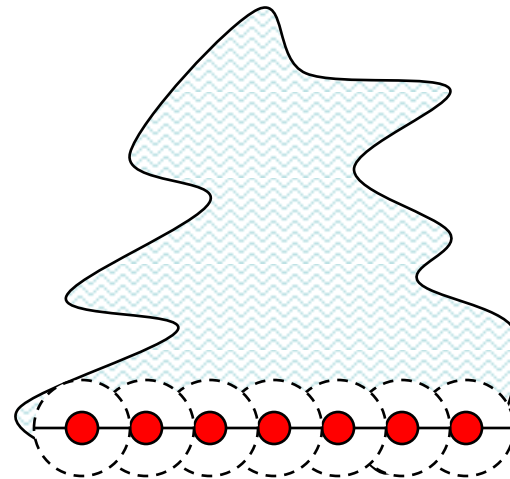
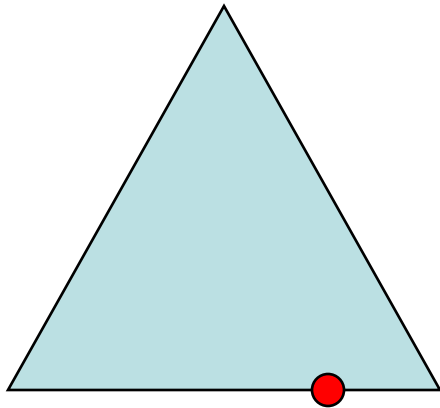
seq. length = n^c

$n = \# \text{ species}$

“very local” metric

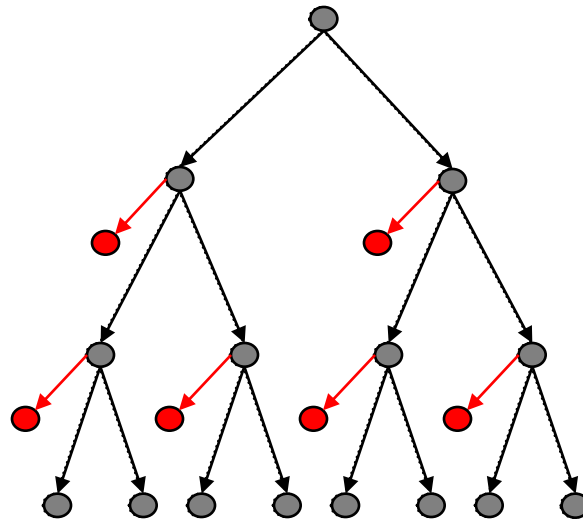
- **recall** - to estimate distances of order M with precision ε , one needs

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



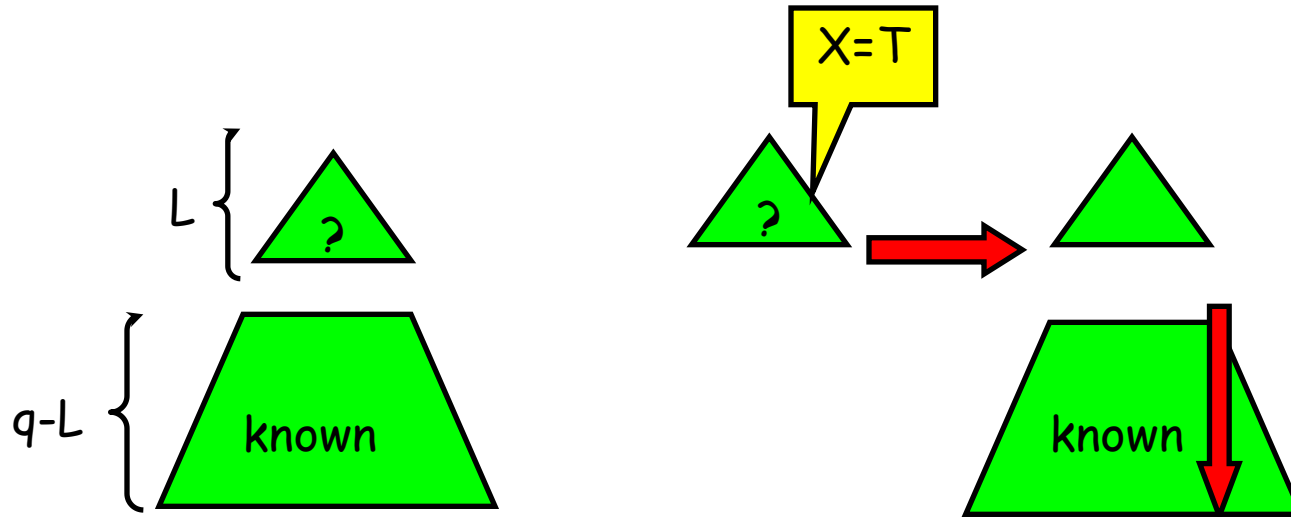
cherry picking algorithm

- loop
 - 1) distance estimation
 - 2) reconstruct one (or a few) level(s)
 - 3) infer sequences at roots



polynomial lower bound

- proof



- **mutual information:** $I(X,Y) = H(X) - H(X | Y)$
- **data processing lemma:** if X and Z are cond. indep. given Y then $I(X,Y) \geq I(X,Z)$

future directions

- empirical testing
 - simulated, real datasets
- understanding classical methods
 - information-theoretic complexity of ML, MP



thank
you