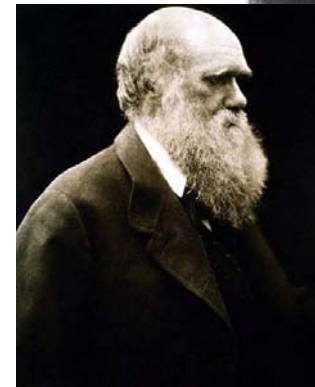
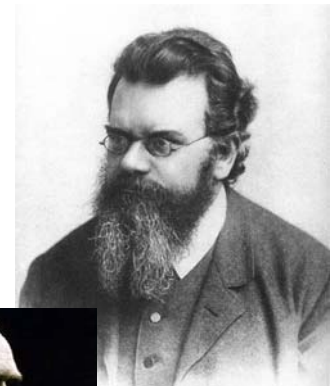


# PHASE TRANSITIONS IN PHYLOGENY

new “optimal” reconstruction algorithms

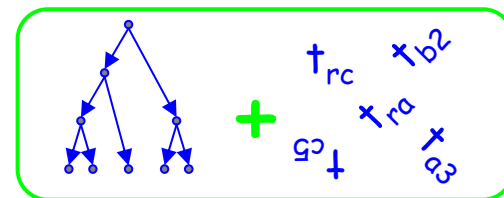
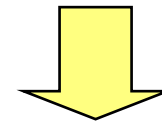
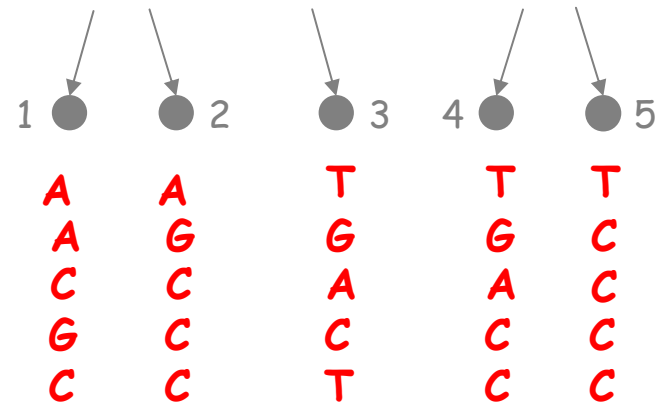
Sebastien Roch  
UC Berkeley

*with:*  
C. Borgs, J. Chayes, C. Daskalakis, E. Mossel



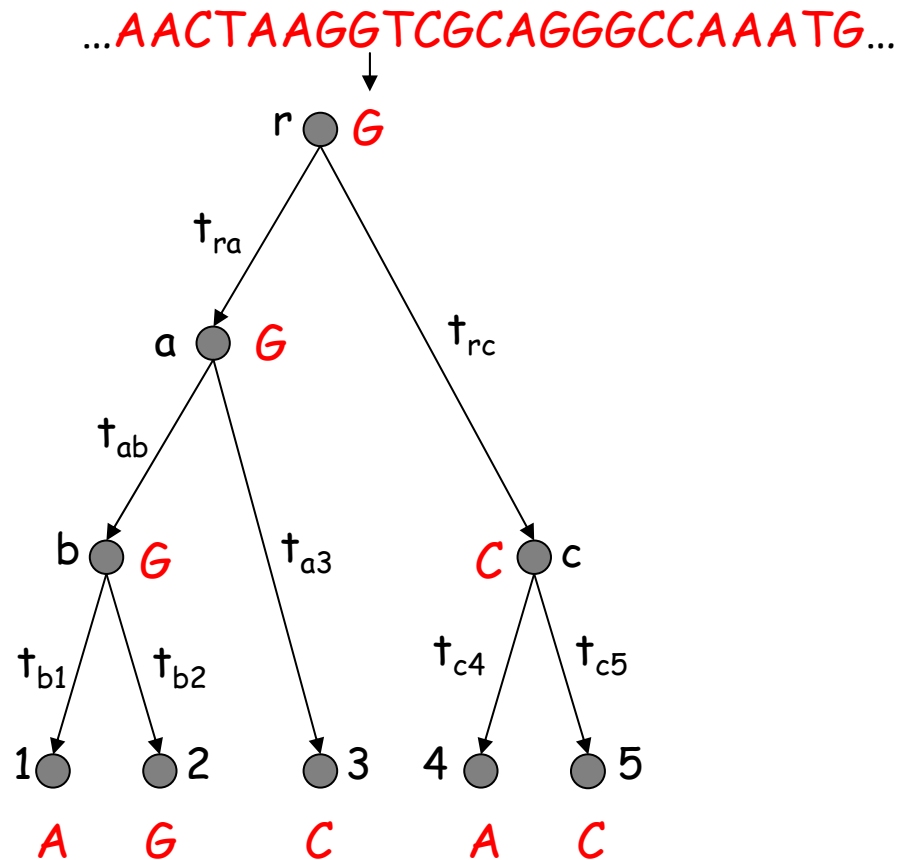
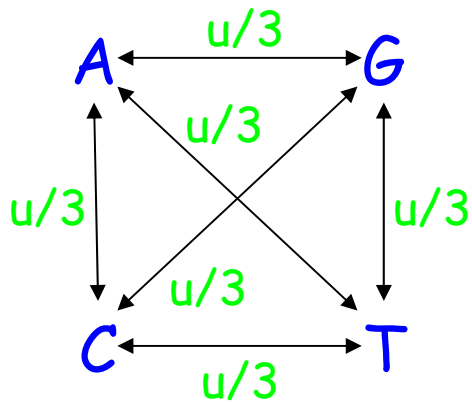
# mathematical analysis of reconstruction

- inferring large-scale phylogenies
- efficiency of reconstruction
  - how long do sequences need to be?
  - how fast is the reconstruction?
- new v. old
  - understanding classical techniques: likelihood, bayesian, parsimony
  - intrinsic limits: new algorithms
- nonstandard applications



# statistical model of evolution

- **Jukes-Cantor model**
  - phylogeny:  $T$
  - rate of mutation:  $u$
  - number of species:  $n$



# maximum likelihood

- likelihood

$$\Lambda(T; D) = \text{Prob}[D | T]$$

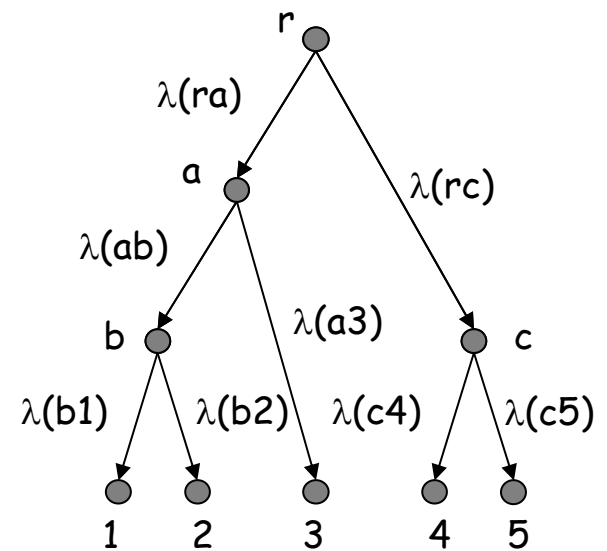
- MLE

$$T^* \leftarrow \max_T \Lambda(T; D)$$

- statistically consistent [Chang'96], but:
  - **result** [Chor-Tuller'05, R'06] - NP-hard (i.e. "computationally intractable"); actually hard to approximate

# distance matrix methods

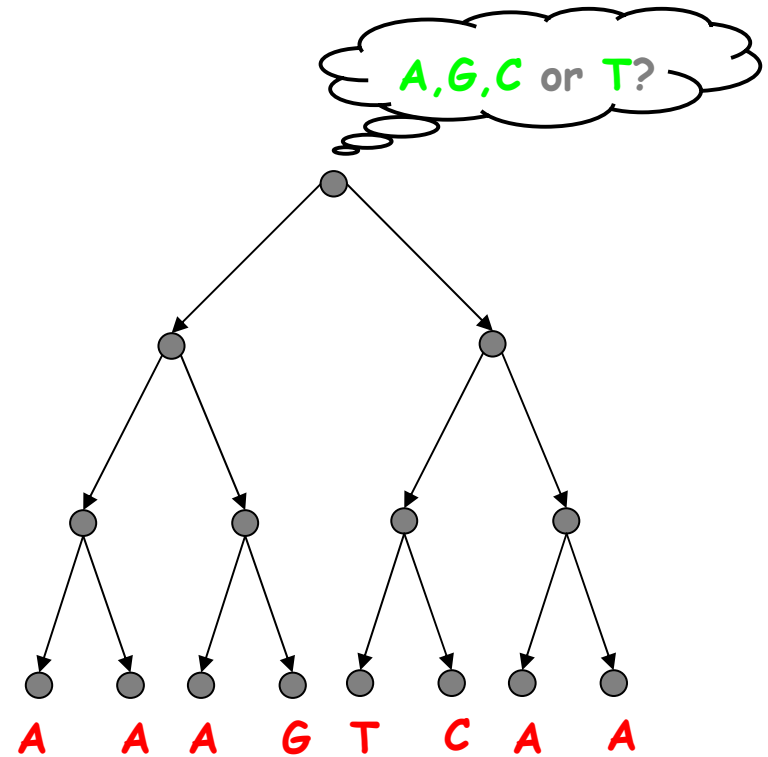
- define a notion of **distance** between leaves [Steel'94]
- reconstruction algorithm:
  - estimate distance from sequences
  - deduce the topology of the tree
- **result** [Erdos-Steel-Szekely-Warnow'97] - sequences of length  $n^c$  suffice
- counting argument: need at least length  $\log n$



$n = \# \text{ species}$

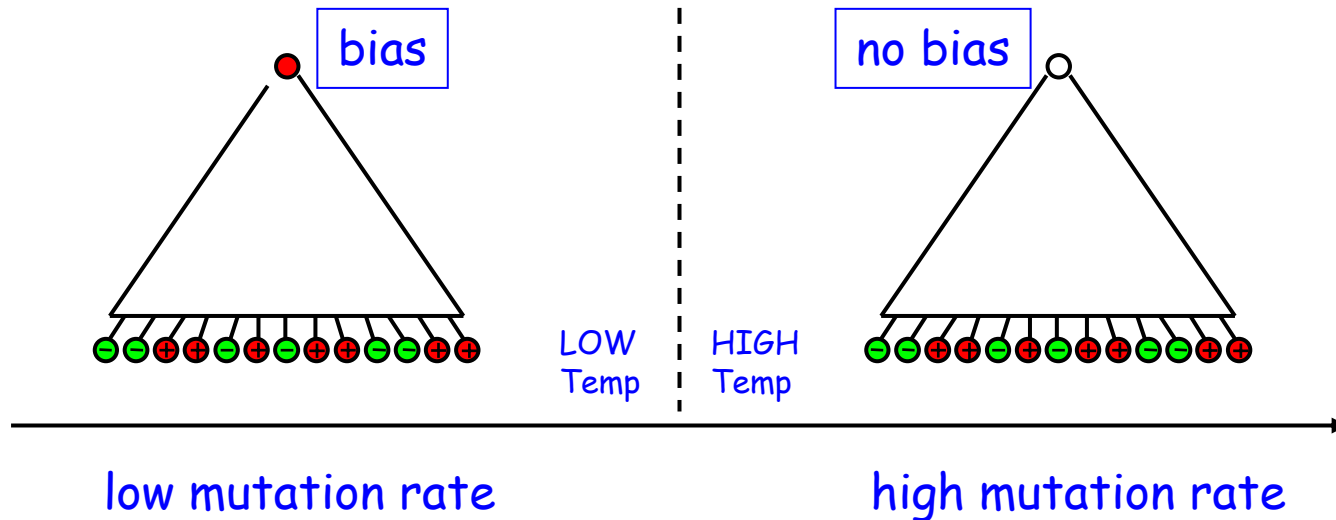
# ancestral reconstruction

- phase transition
  - trade-off between noise and duplication

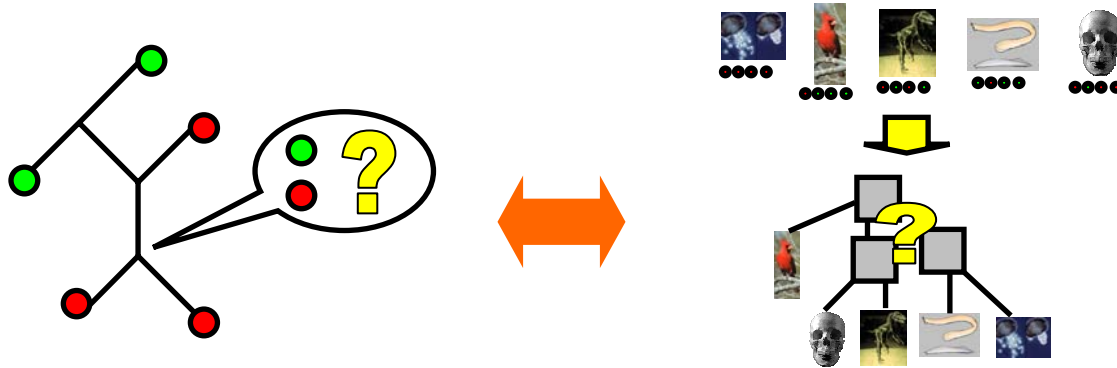


# connection to statistical physics

- **result** - sharp transition at critical mutation rate
  - [Bleher-Ruiz-Zagrebnoy'95], [Ioffe'96], [Evans-Kenyon-Peres-Schulman'00], [Kenyon-Mossel-Peres'01], [Martinelli-Sinclair-Weitz'04], [Borgs-Chayes-Mossel-R'06], [Higuchi'77], [Kesten-Stigum'66]



# resolution of Steel's conjecture



ancestral  
reconstruction

phylogenetic  
reconstruction

[Daskalakis-  
Mossel-R'06]

low mutation rate



seq. length =  $\log n$

[Mossel'04]

high mutation rate

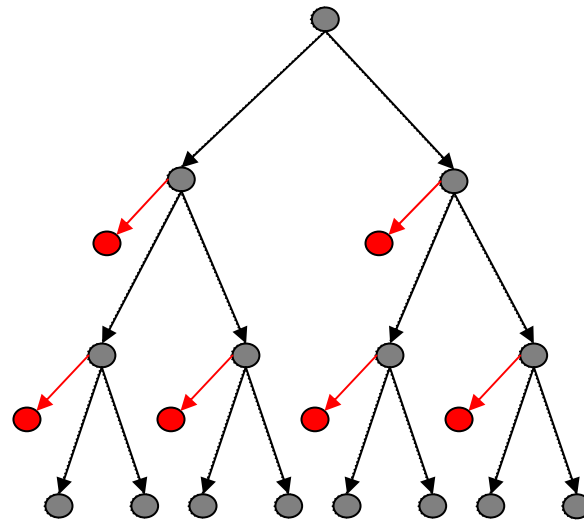


seq. length =  $n^c$

$n = \# \text{ species}$

# basic reconstruction algorithm

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



# network tomography

- **multicast delay inference**

[Castro et al.'03]

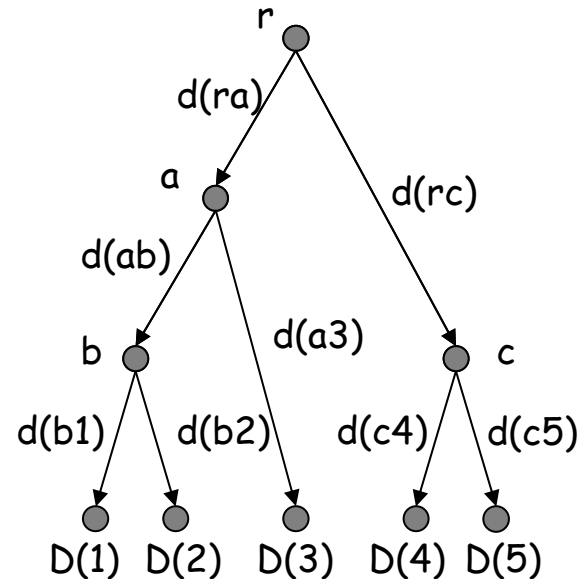
- routing tree
- root sends a packet to all leaves
- delay on each link

- **problem**

- using measures of delays at the leaves, recover tree topology and delay distributions

- **idea** [Bhamidi-Rajagopal-R'06]

- construct additive metrics from delays
- apply distance methods



$$D(i) = \sum_{u \in P(r,i)} d(u)$$

thank  
you

