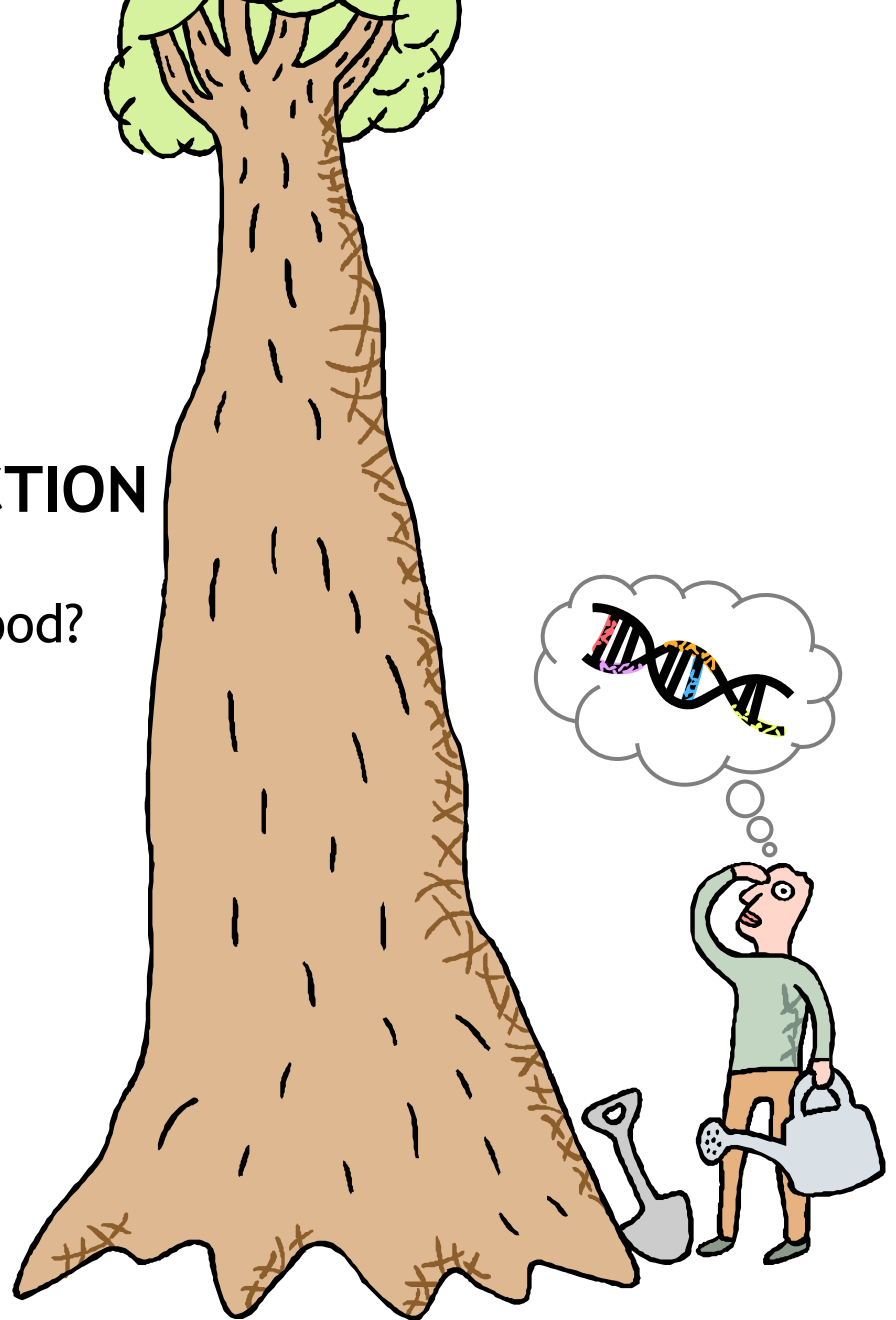


# PHYLOGENY RECONSTRUCTION

are distance-matrix methods  
as accurate as maximum likelihood?

Sebastien Roch  
UCLA

UCLA Bioinformatics Seminar  
November 23, 2009



# outline of the talk

PART 0  
review:  
distance-matrix methods



PART I  
insights from  
statistical physics



PART II  
beyond the  
oracle view



PART 0  
review: distance-matrix  
methods



PART I  
insights from  
statistical physics

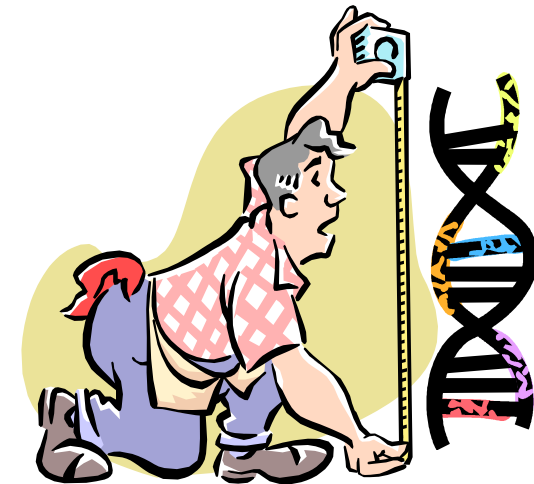


PART II  
beyond the  
oracle view



# PART 0

## background: how good are distance methods?





# stochastic model of evolution

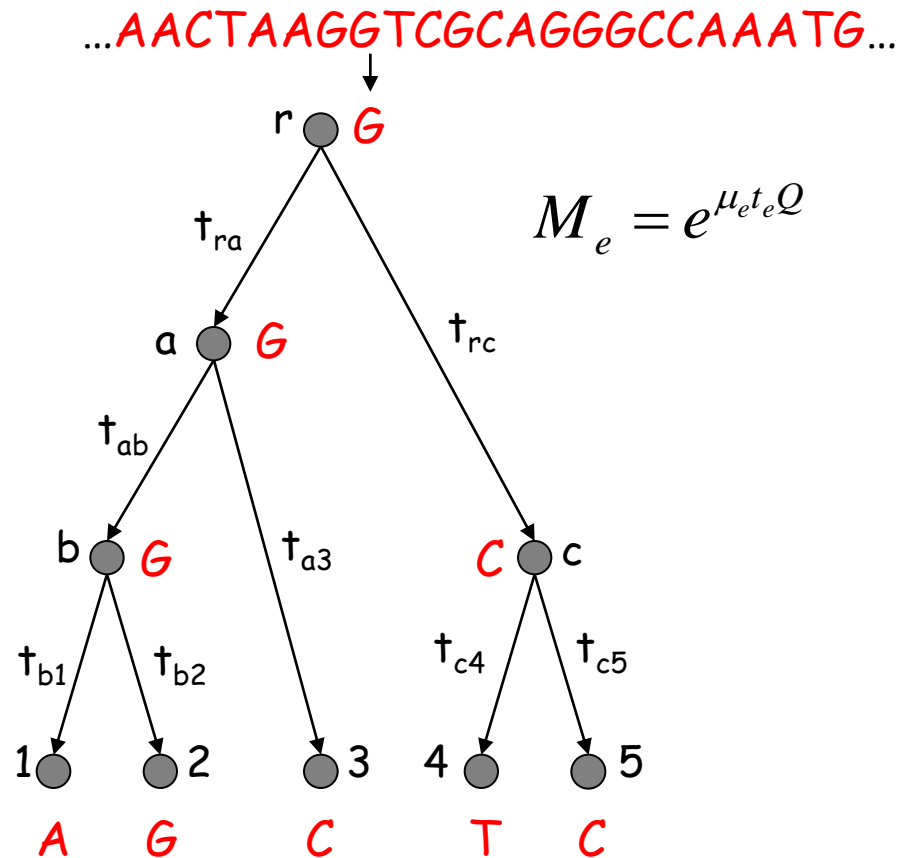
- **Jukes-Cantor model/Potts model with free boundary**

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

$$Q = \left( \begin{array}{cc} A & G \\ C & T \end{array} \right)$$

- **remark**

- no deletion/insertion



# phylogeny reconstruction

- **setup**

- sequence  $s_a^1, \dots, s_a^k$  for each species
- trees on  $n$  leaves:  $T_n$
- estimator:

$$\Psi_n : \left\{ \left( s_a^i \right)_{i=1}^k \right\}_{a \in L} \mapsto T \in T_n$$

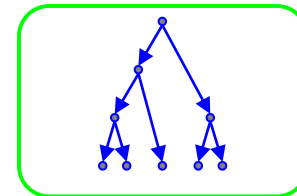
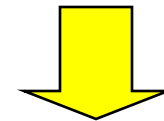
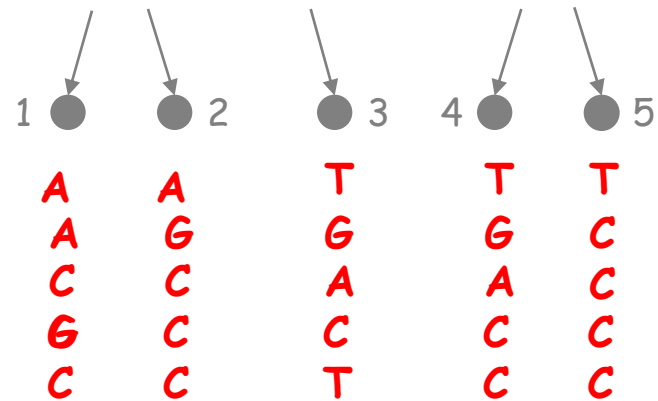
- how to **compare** different methods?

- **computational efficiency**

- **consistency** -

$$P[\text{correct reconstruction}] \rightarrow 1$$

as the sequence length goes to infinity

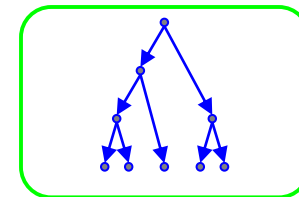
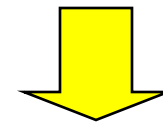
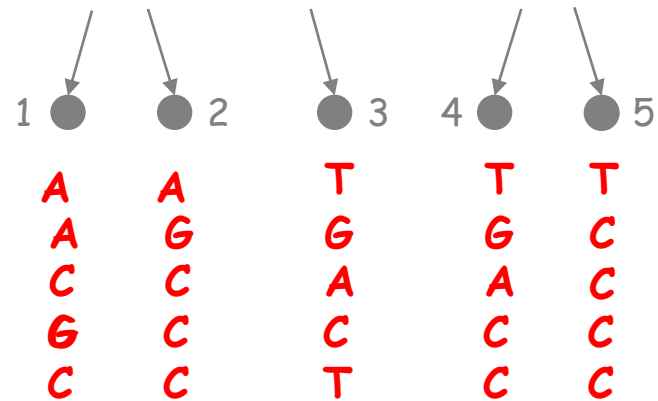


# asymptotic sample complexity

- **definition** - the estimator  $\Psi_n$  solves the **phylogenetic reconstruction problem** with  $k$  samples and confidence  $1-\delta$  if

$$P\left[\Psi_n\left(\left\{\left(s_a^i\right)_{i=1}^k\right\}_{a \in L}\right)=T\right] \geq 1-\delta$$

- **asymptotic sample complexity** - how does  $k$  **scale** as a function of
  - $n$ : number of terminal taxa
  - $f$ : shortest branch length
  - depth of the tree
- **goal** - compare how well different methods **extract** phylogenetic signal

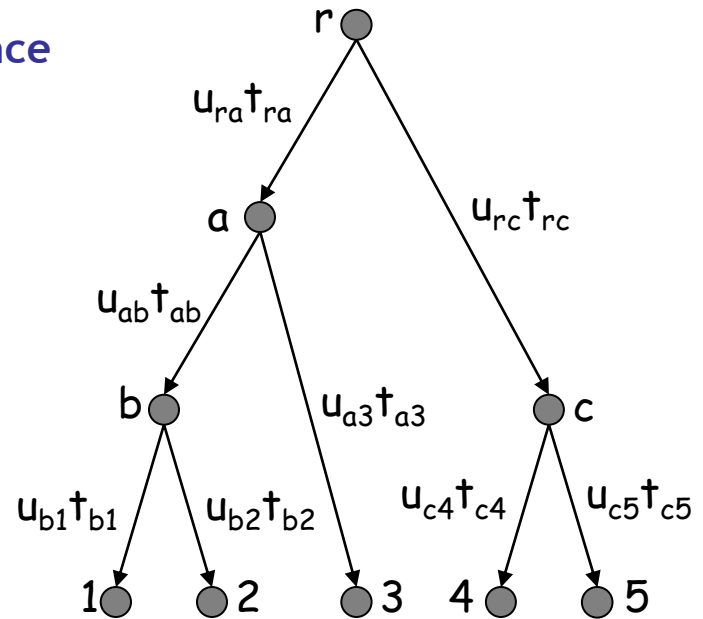


# distance-matrix methods

- 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 matrix

- **data:** n aligned sequences

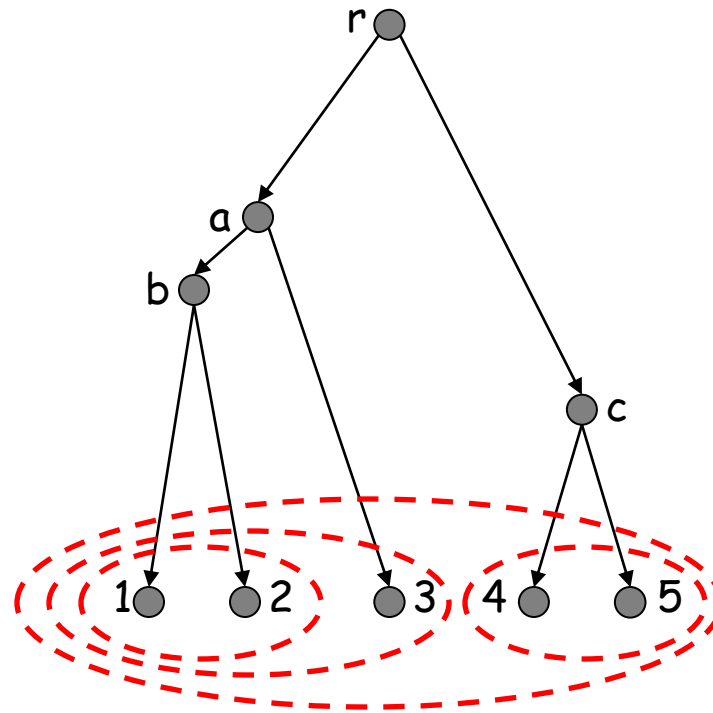
	1	2	3	4	5	6	7	8	9	0	1	
Homo sapiens	A	-	C	A	T	G	G	A	G	-	A	A
Pan	A	-	T	A	A	T	A	-	A	G	C	A
Gorilla	A	T	C	A	-	C	A	-	A	G	C	G

- $p'(i,j)$ : **proportion** of homologous sites that disagree between sequences  $i$  and  $j$  (ignoring columns with gaps)
  - example:  $p'(\text{Homo sapiens}, \text{Pan}) = 0.2$
- **Jukes-Cantor formula**

$$D'(i, j) = -\frac{3}{4} \log \left( 1 - \frac{4}{3} p'(i, j) \right)$$

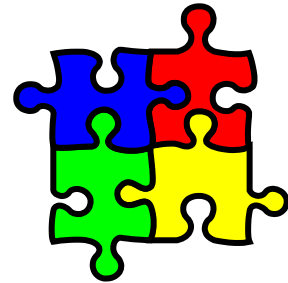
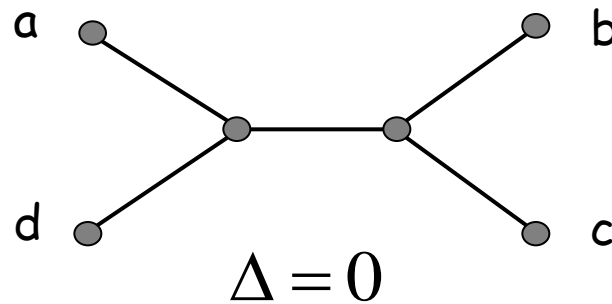
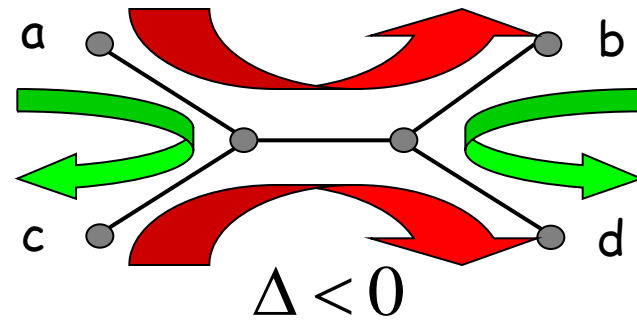
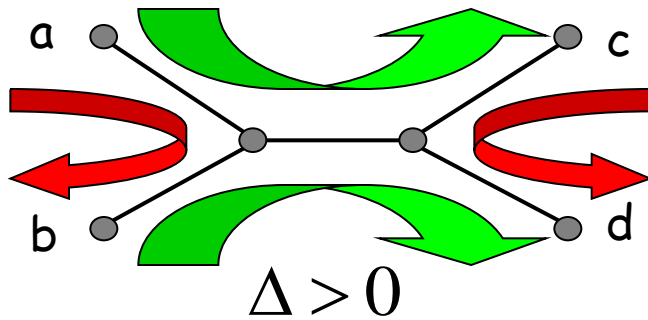
# molecular clock: agglomerative algorithm

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



# beyond the molecular clock

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



# classical ASC result

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

- **theorem [ESSW'99]** - UPGMA and Short Quartet Method have ASC

$$k \propto \frac{1}{f^2} \cdot \log n \cdot \exp(\text{Depth})$$

- **proof idea** - recall

$$D'(i, j) = -\frac{3}{4} \log \left( 1 - \frac{4}{3} p'(i, j) \right)$$

by CLT, to estimate a distance of order  $M$  with precision  $\varepsilon$ , one needs

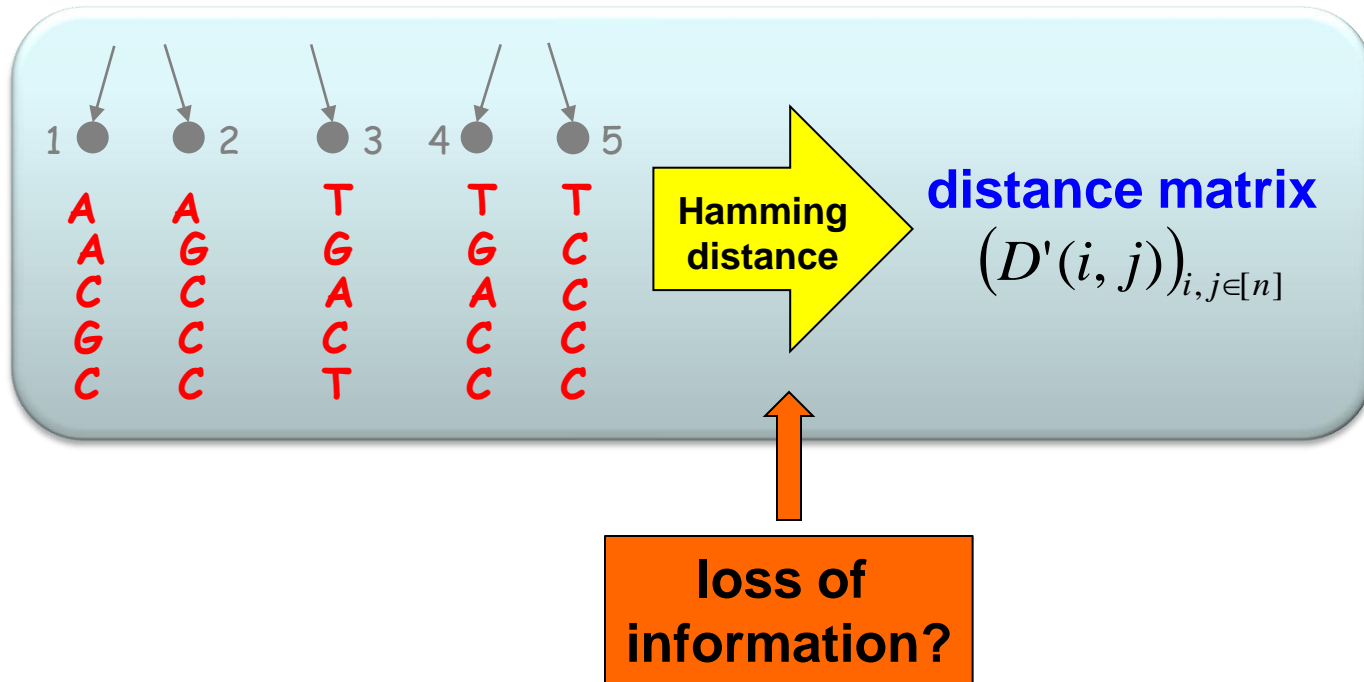
$$k \propto \frac{1}{\varepsilon^2} \exp(M)$$

note: the  $\log n$  factor comes from a Bonferroni correction

# distance-based v. likelihood-based

- **theorem [ESSW'99]** - UPGMA and Short Quartet Method have ASC

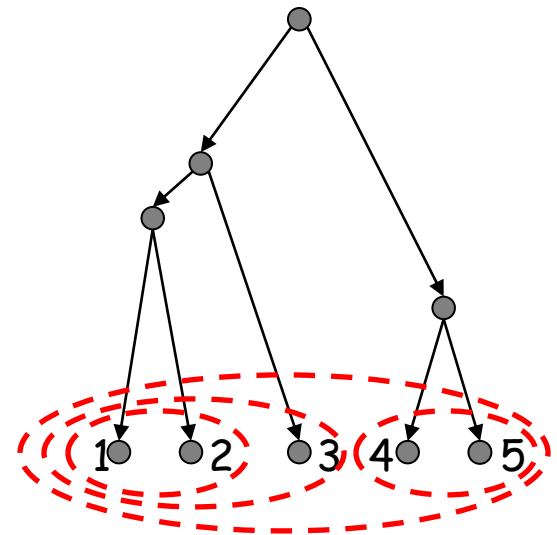
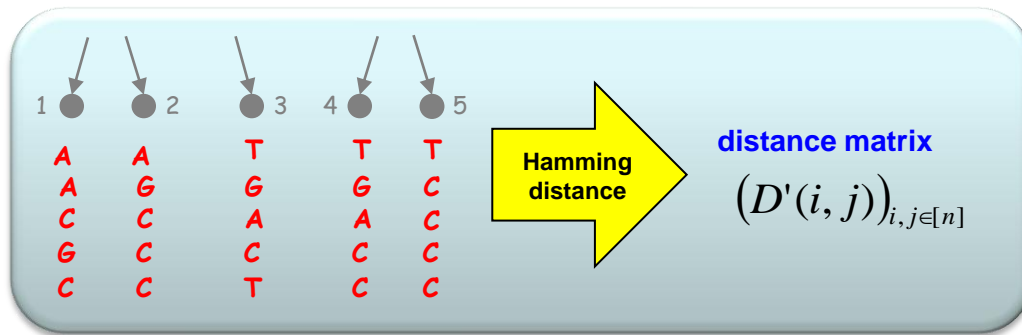
$$k \propto \frac{1}{f^2} \cdot \log n \cdot \exp(\text{Depth})$$



# main result

- assume  $0 < f < t(e) < g < g^*$ , for all  $e$
- **theorem [R'09]** - there is a distance-based method that only requires

$$k \propto \frac{1}{f^2} \cdot \log n$$



PART 0  
review: distance-matrix  
methods



PART I  
insights from  
statistical physics

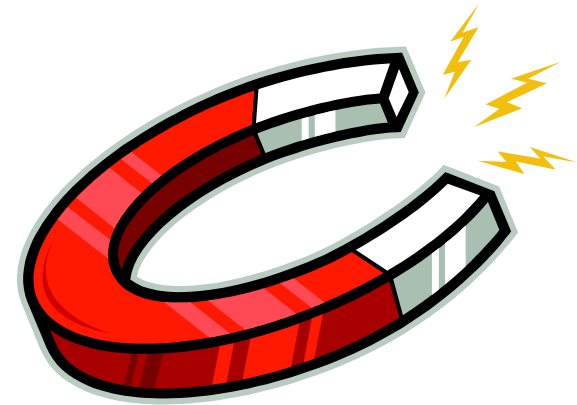


PART II  
beyond the  
oracle view



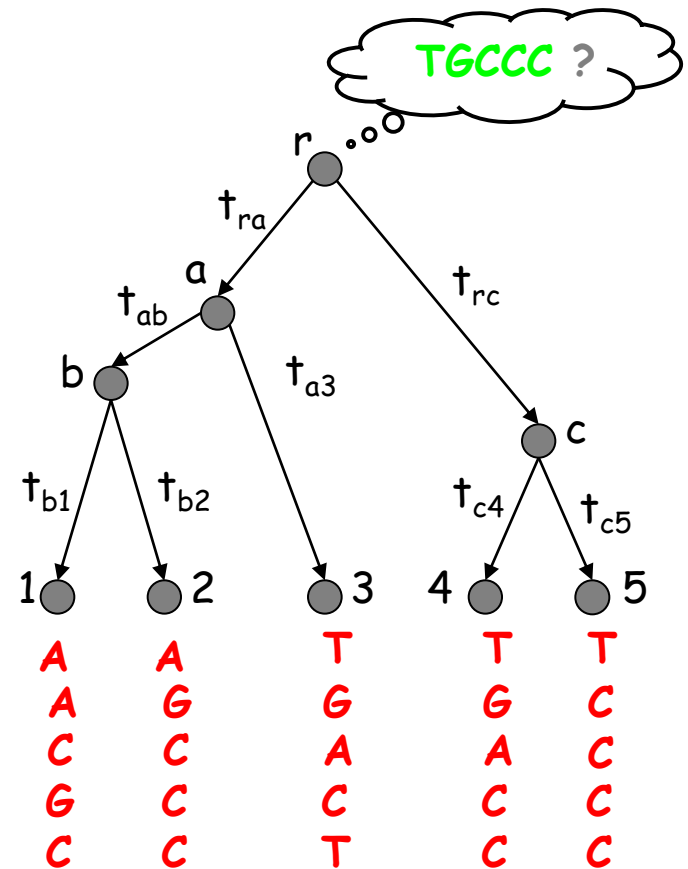
# PART I

## 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



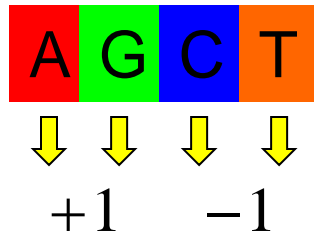
# phase transition

- Kesten-Stigum bound [Kesten-Stigum'67, Evans et al.'00, Mossel-Peres'03]

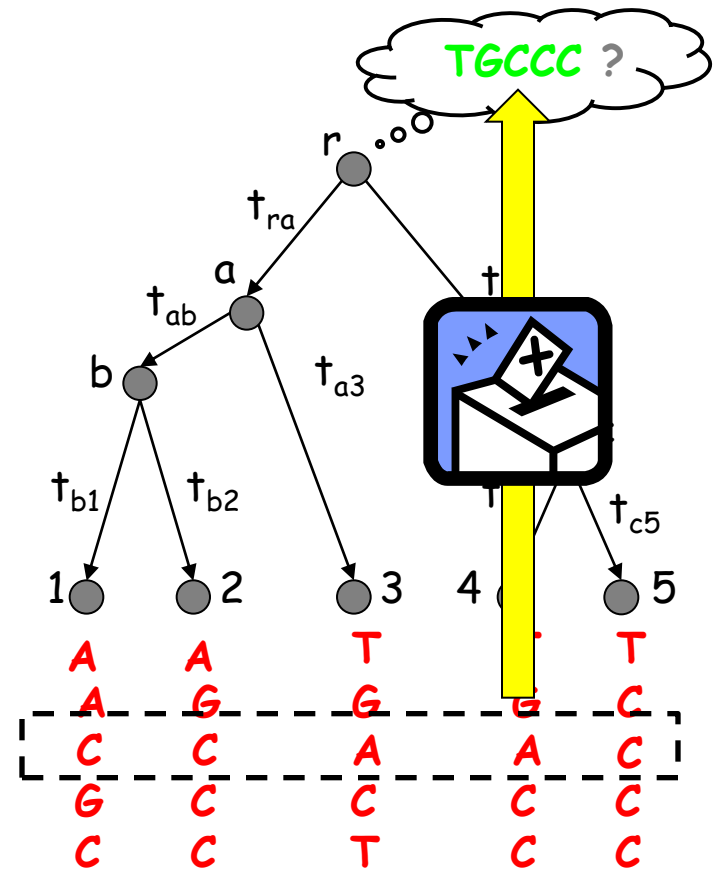
- bound on critical branch length:

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

- root estimator:

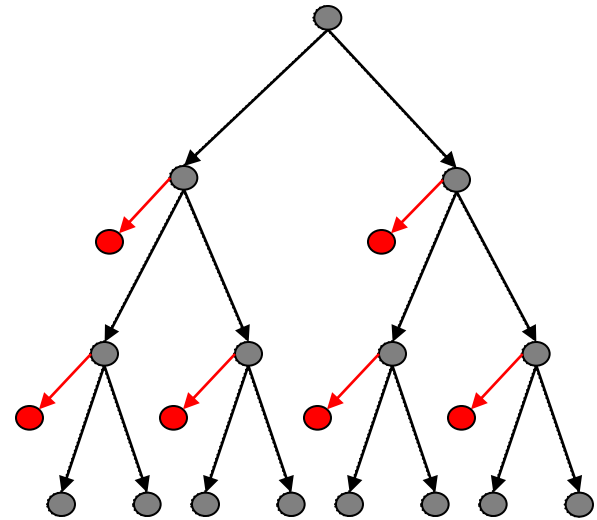


$$S = \sum_{a \in [n]} 2^{-|a|} s_a^i$$



# “boosted” algorithm

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



ancestral  
reconstruction

phylogenetic  
reconstruction

[Daskalakis-  
Mossel-R]

reconstruction



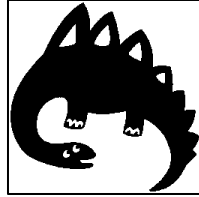
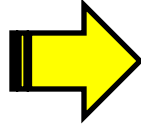
$$k \propto f^{-2} \cdot \log n$$

[Mossel'04]

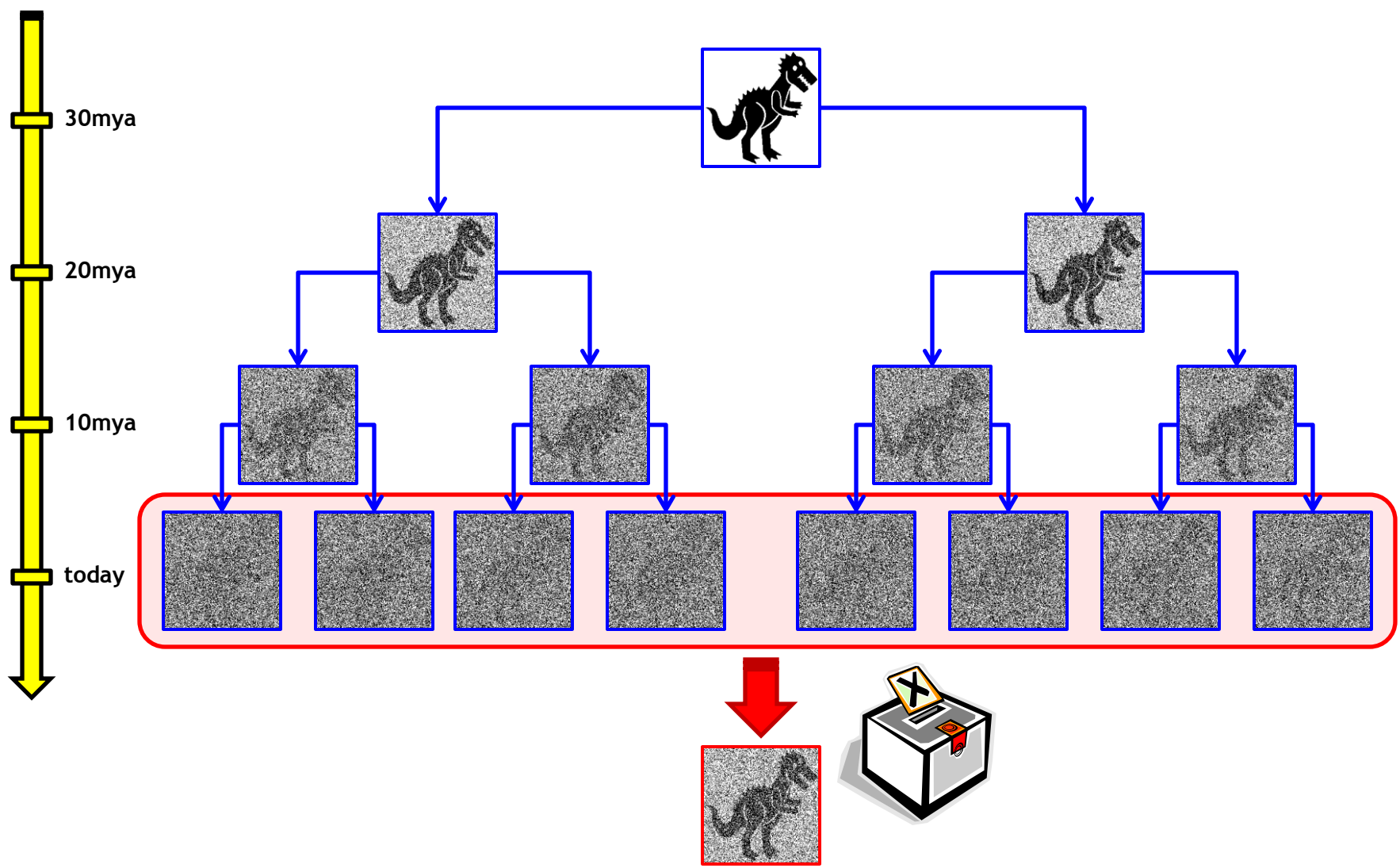
non-reconstruction



$$k \propto f^{-2} \cdot \log n \cdot e^{\text{Depth}}$$









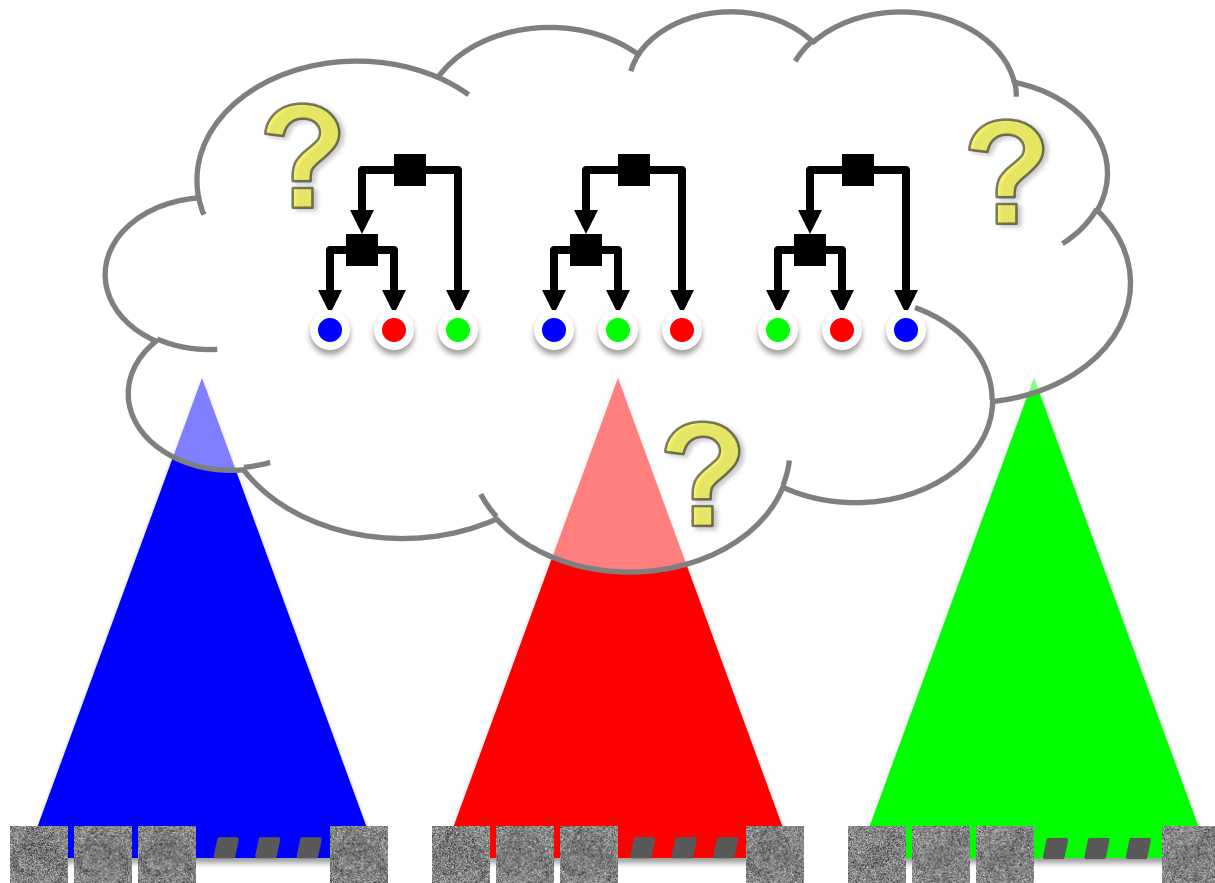
40mya

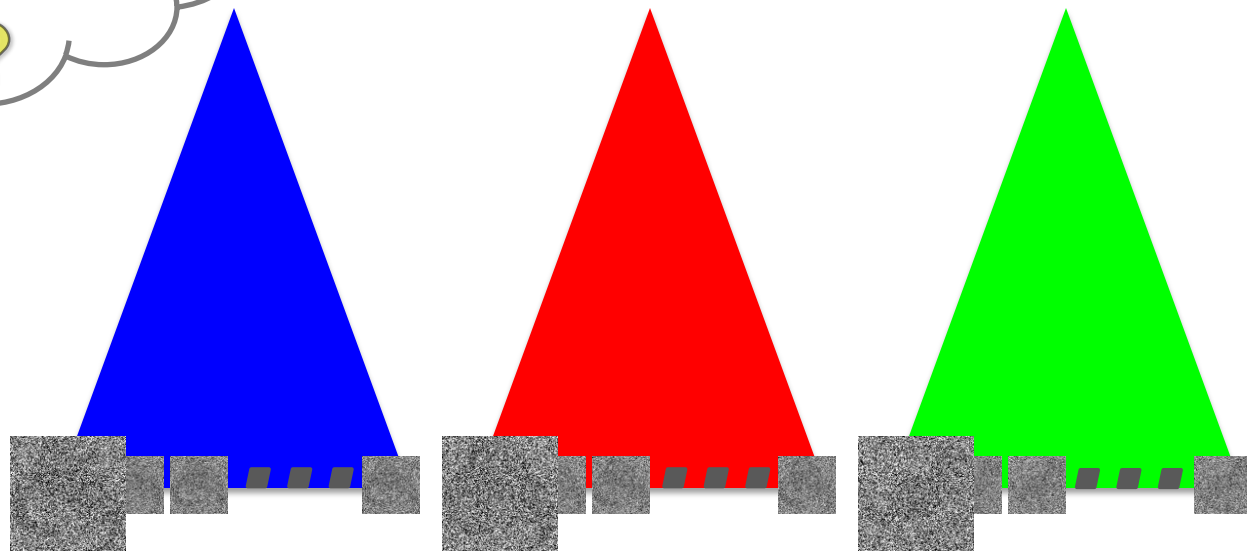
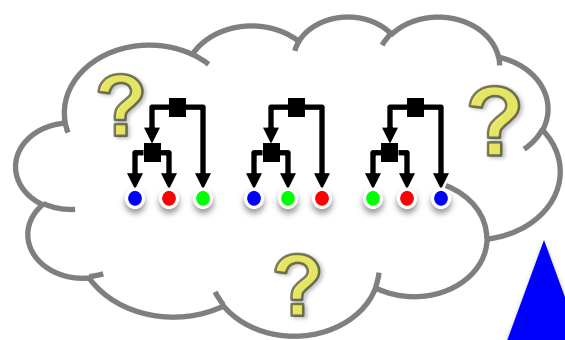
30mya

20mya

10mya

today



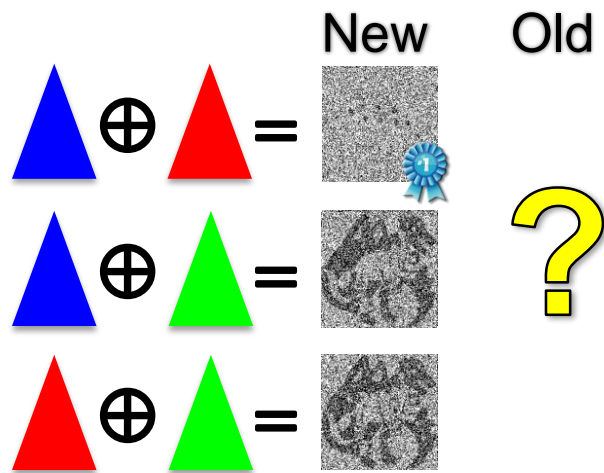
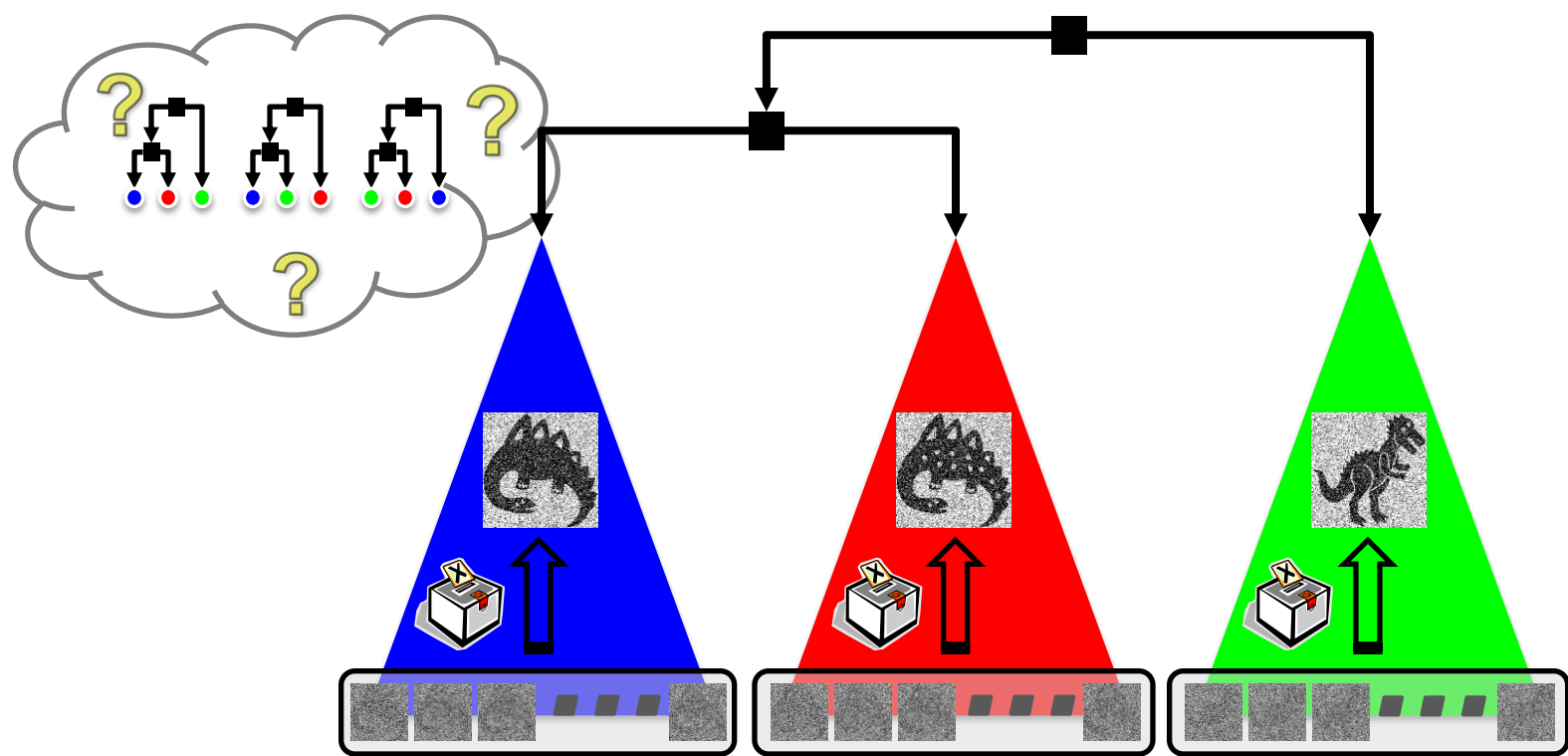


$$\triangle_{\text{blue}} \oplus \triangle_{\text{red}} = \text{gray block}$$

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

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

?



PART 0  
review: distance-matrix  
methods



PART I  
insights from  
statistical physics

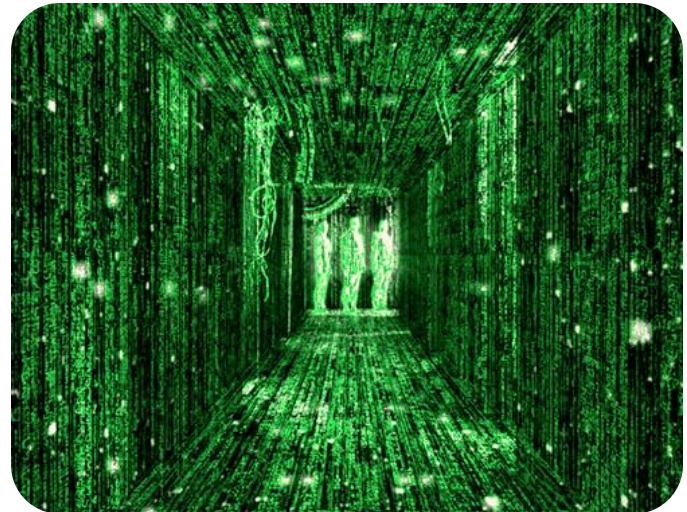


PART II  
beyond the  
oracle view



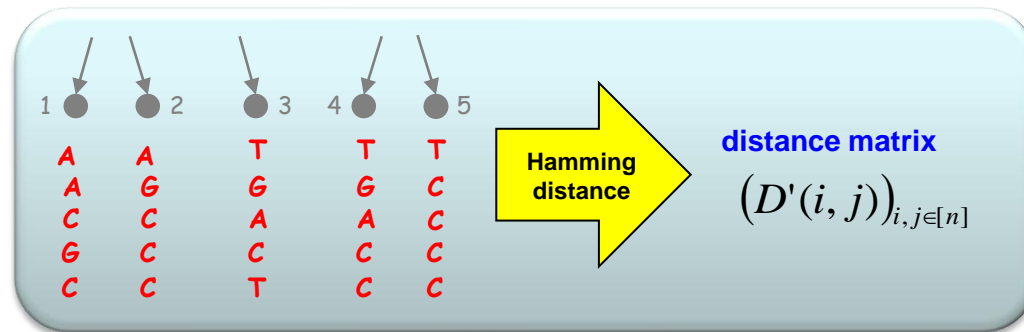
# PART II

## what really lies behind the matrix?



# can distance methods achieve $f^{-2}\log n$ ?

- “obvious” answer: no
  - mapping from dataset to distance matrix is far from invertible
  - only infer “pairwise” distributions
  - how to reconstruct ancestral sequences from aggregated information?
  - oracle view...



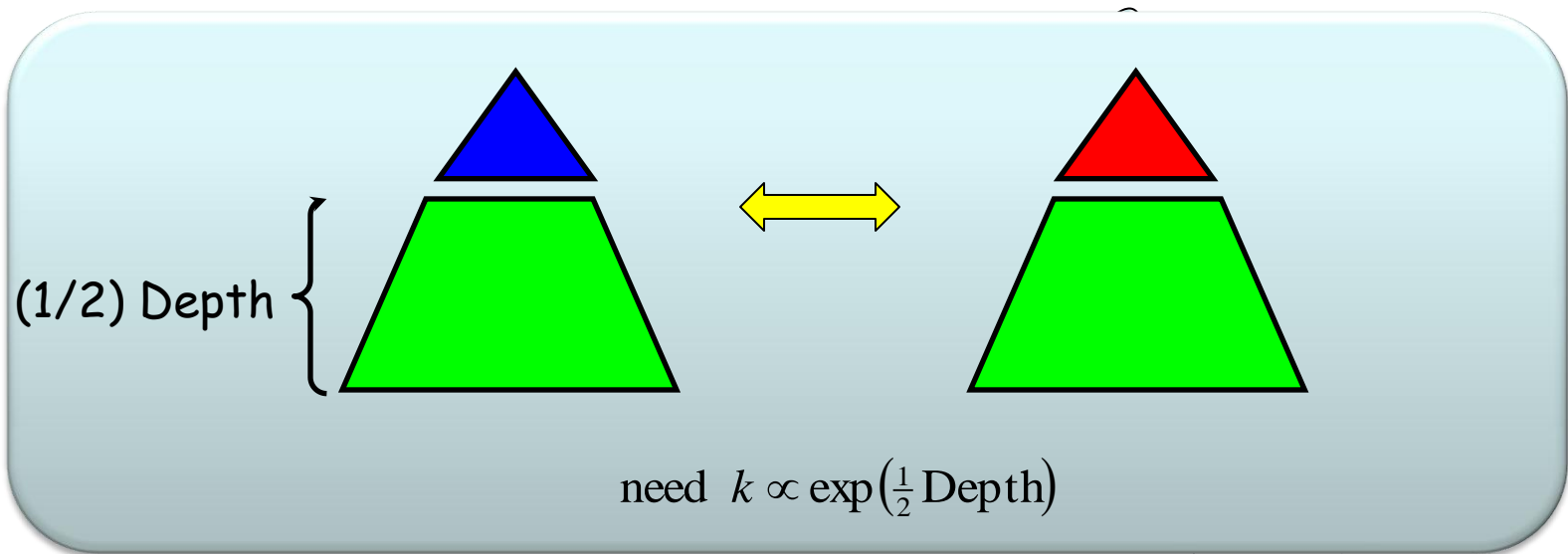
# oracle view of the distance matrix

- **fact** - to estimate all distances of order  $M$  with precision  $\varepsilon$ , one needs

$$k \propto \frac{1}{\varepsilon^2} \cdot \log n \cdot \exp(M)$$

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

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



# the matrix reloaded

- **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  $(a,b,c,d)$

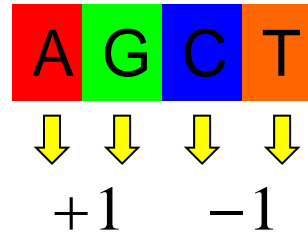
$$\mu_{\{a,b,c,d\}}$$

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



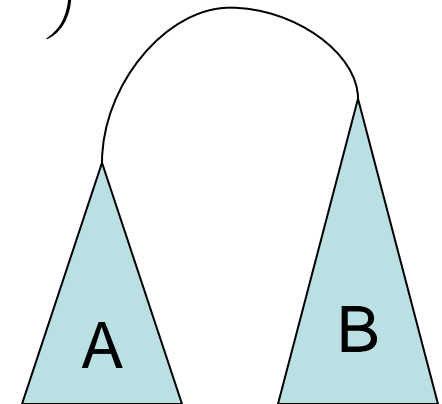
# revisiting the averaging procedure I

- **step 1** - project the states to binary values



- the distance matrix becomes

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



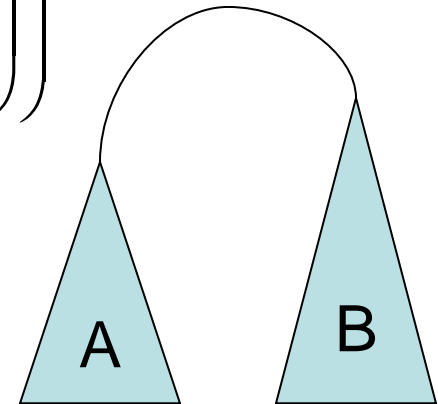
# revisiting the averaging procedure II

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

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

“majority”

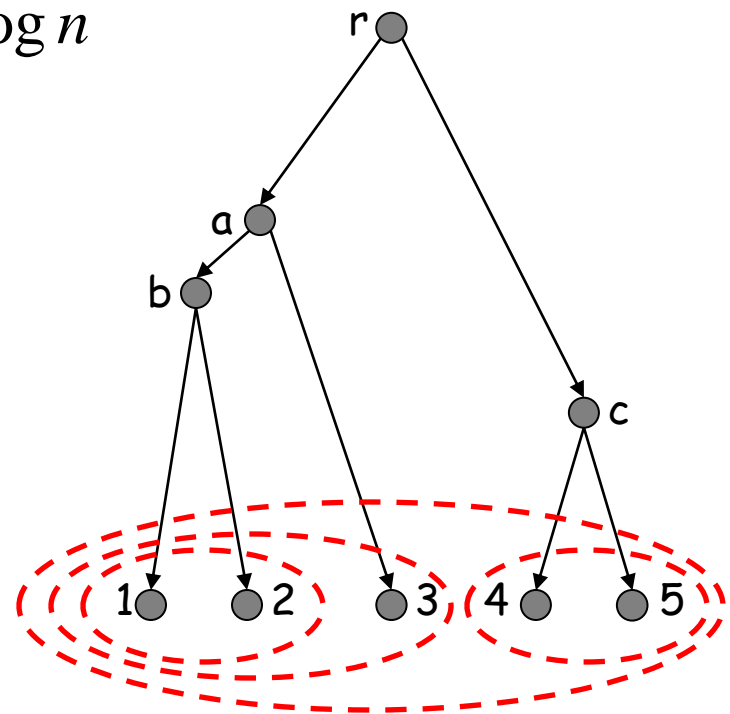
$$D'(a, b) = -\ln \left( \frac{1}{k} \sum_{i=1}^k s_a^i s_b^i \right)$$



# sample complexity of distance methods

- assume  $0 < f < t(e) < g < g^*$ , for all  $e$
- **theorem [R'09]** - clustering with “exponential” averaging only requires

$$k \propto \frac{1}{f^2} \cdot \log n$$



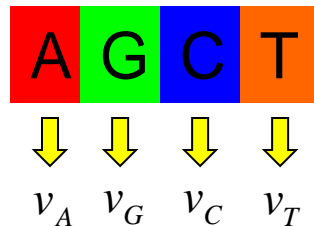
# on the importance of averaging

- **practical implications** - must do distance averaging to extract all the information from the matrix (e.g. quartet-based approaches do not), BUT...
  - use the right averaging **weights** (UPGMA v. WPGMA)

$$D'(A \cup B, C) = \frac{|A|}{|A \cup B|} D'(A, C) + \frac{|B|}{|A \cup B|} D'(B, C)$$

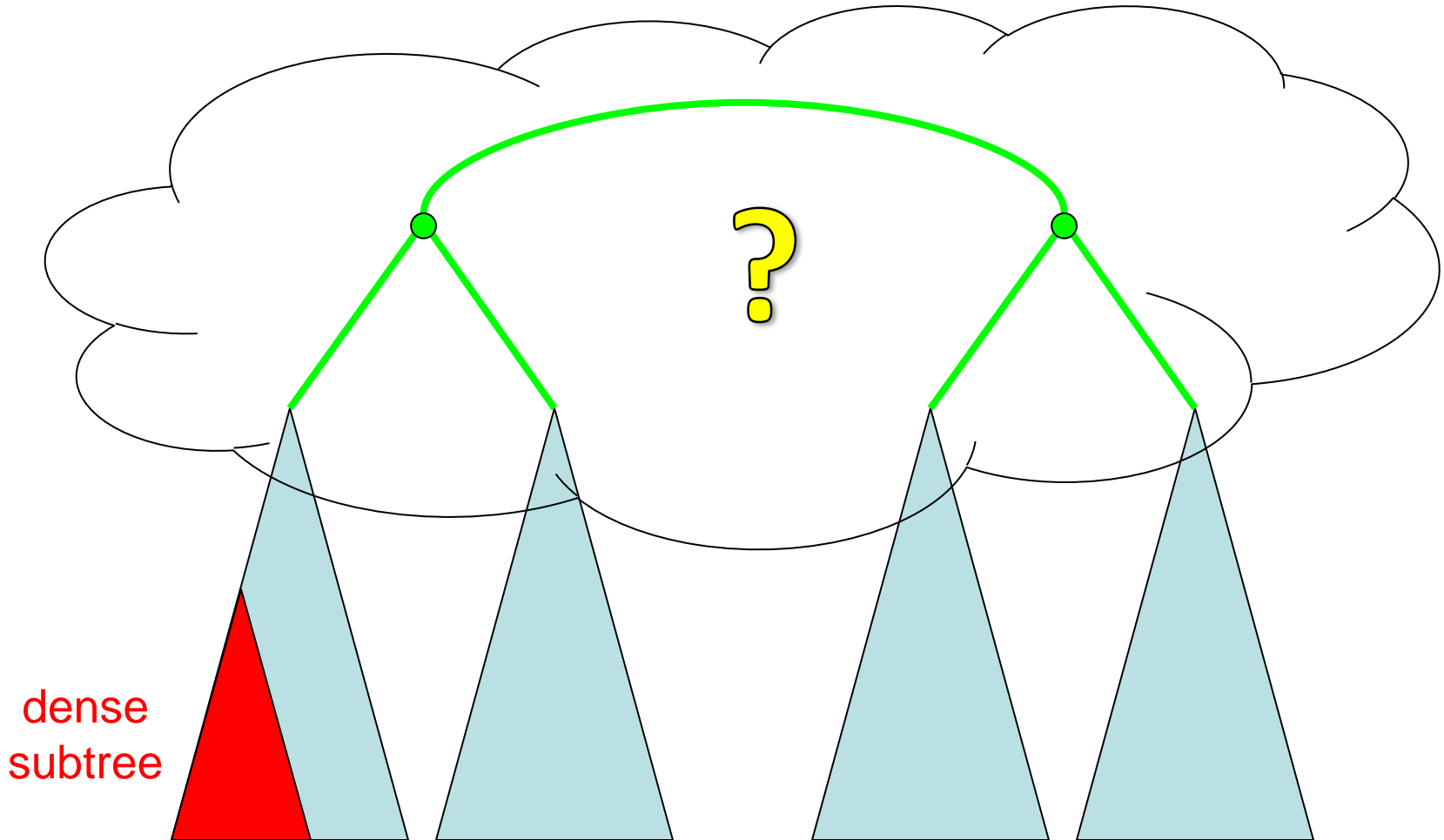
$$D'(A \cup B, C) = \frac{1}{2} D'(A, C) + \frac{1}{2} D'(B, C)$$

- use **uncorrected** distances (but correct before a clustering step)
- use the right **estimate** of distance (logdet distance may not be best)

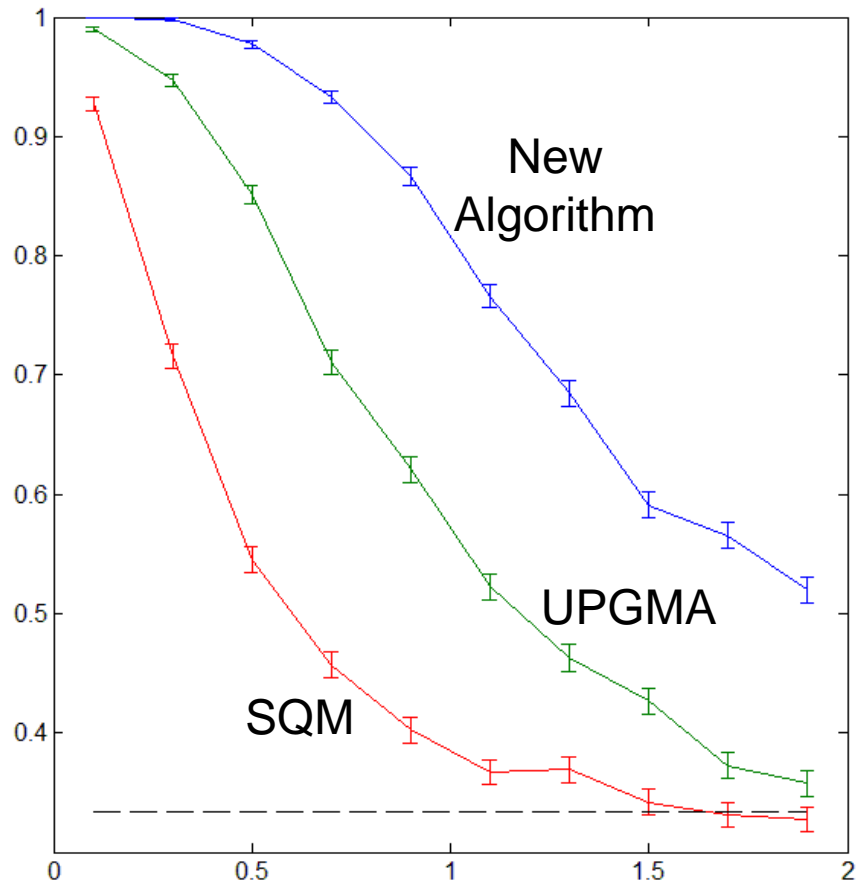


$$D'(a, b) = -\ln \left( \frac{1}{k} \sum_{i=1}^k s_a^i s_b^i \right)$$

# controlled experiment: setup

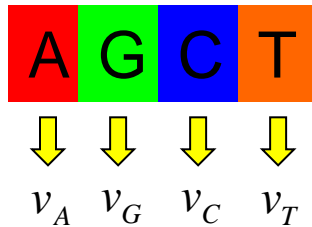


# controlled experiment: results



# more results I

- **general time-reversible models** - consider a reversible rate matrix  $Q$
- **Kesten-Stigum** - the right estimator of ancestral sequences involves the second right eigenvector  $v$  of  $Q$
- **new distance estimator** - map states to values of  $v$



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

# more results II

- **general trees** - consider general tree with uneven mutation rates
- **Kesten-Stigum** - have to use **weighted** majority. the weights are related to the path lengths
- **issues** - have to infer path lengths as well. plus, more complicated combinatorial algorithm.

# acknowledgments

- Elchanan Mossel (UC Berkeley)
- Yuval Peres (Microsoft)
- Constantinos Daskalakis (MIT)

**thank**  
**you**