

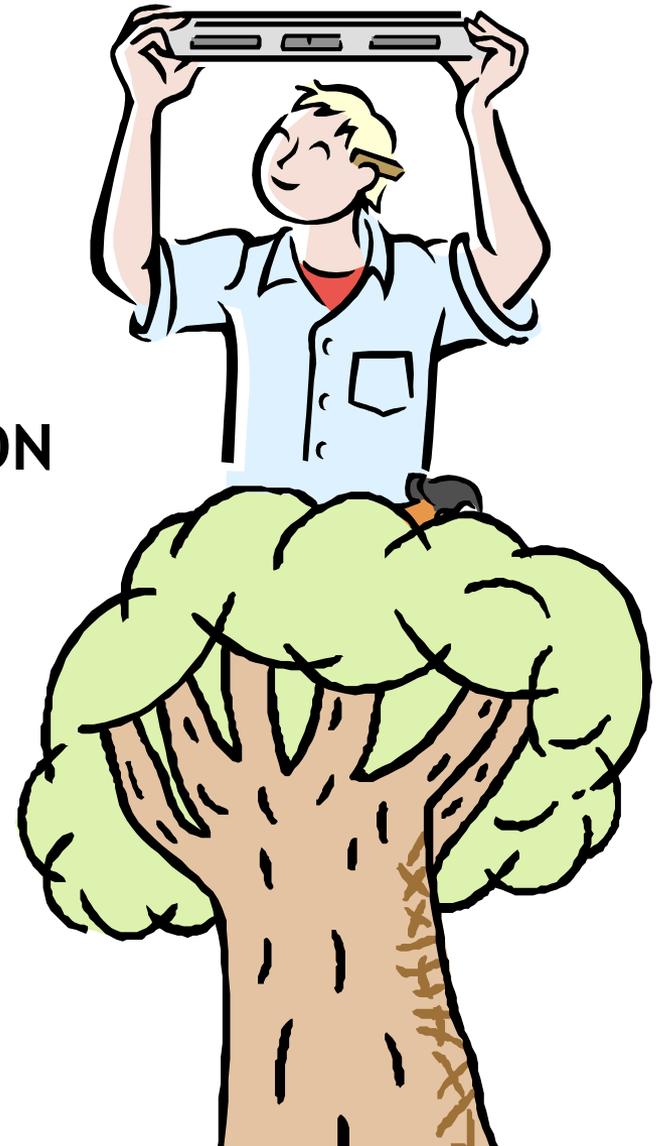


ALIGNMENT-FREE PHYLOGENETIC RECONSTRUCTION

Sebastien Roch (UW-Madison)

with:

Constantinos Daskalakis (MIT)



theoretical analysis of phylogenetic reconstruction methods

- **setup**

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

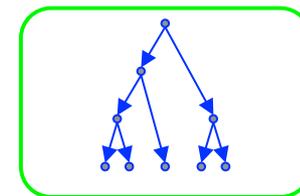
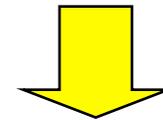
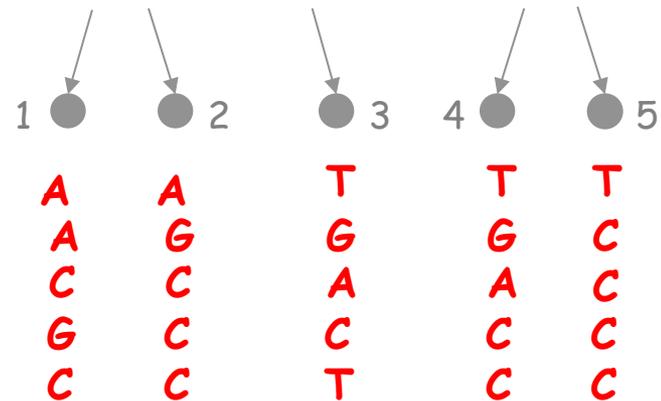
$$\Psi_n : \left\{ \left(s_a^i \right)_{i=1}^k \right\}_{a \in L} \mapsto T \in T_n$$

- how to **compare** different methods?

- **computational efficiency**
- **statistical consistency (& rate)**

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

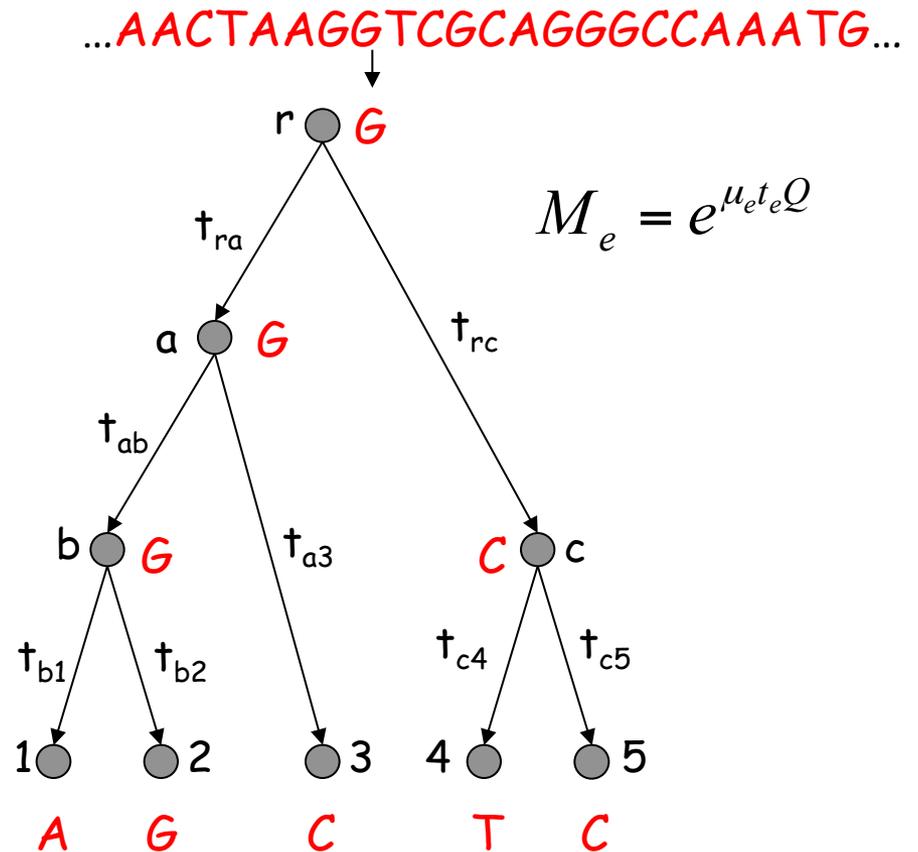
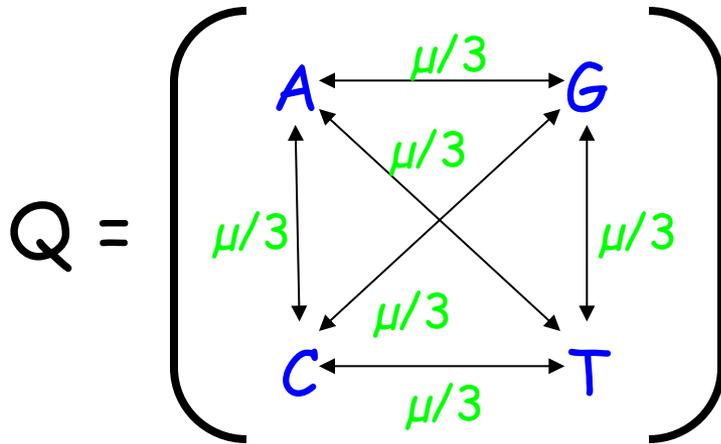
as the sequence length goes to infinity under a statistical model of sequence evolution



substitution-only model of sequence evolution

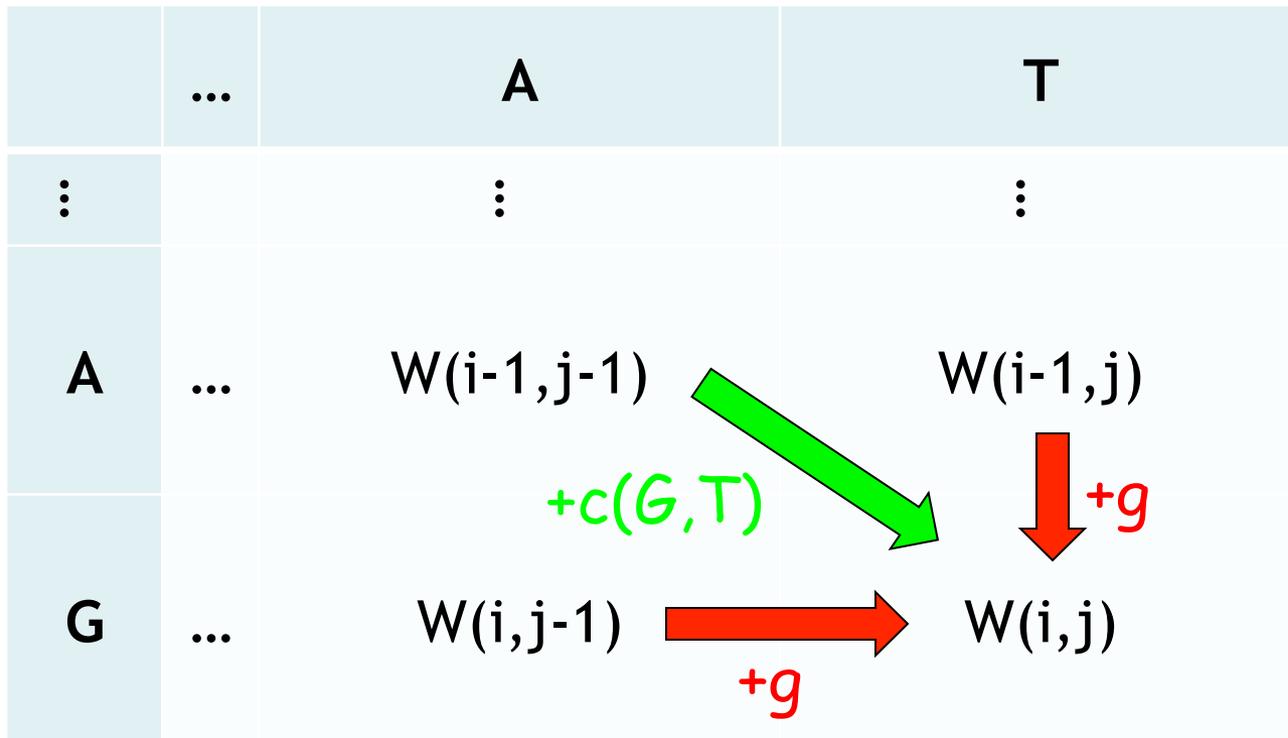
- **Example: Jukes-Cantor model**

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



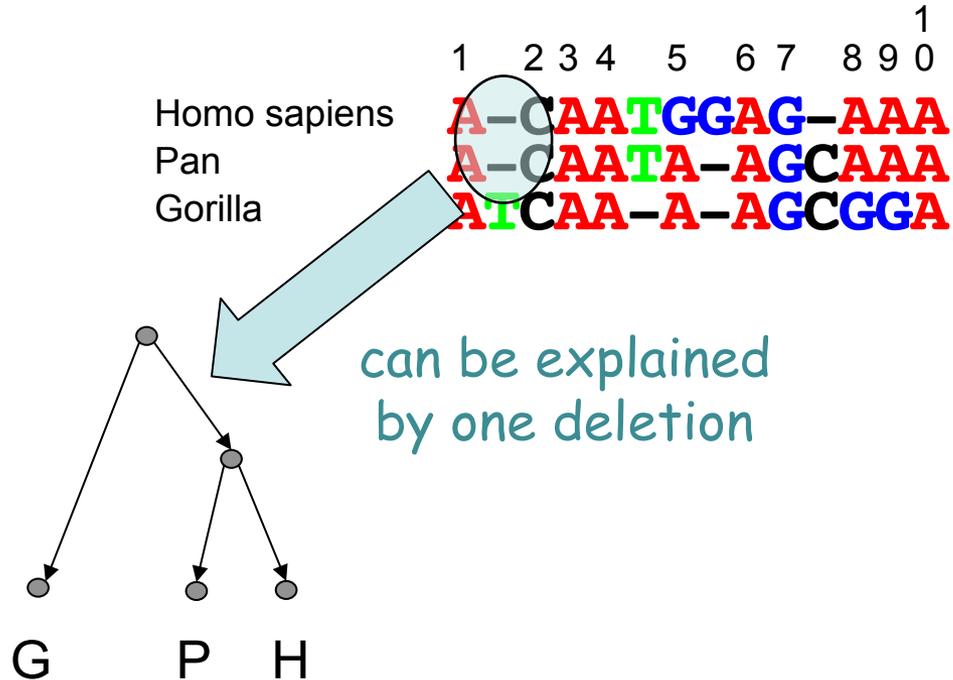
issues with alignment

- **dimensionality curse** : takes time $O(k^n)$

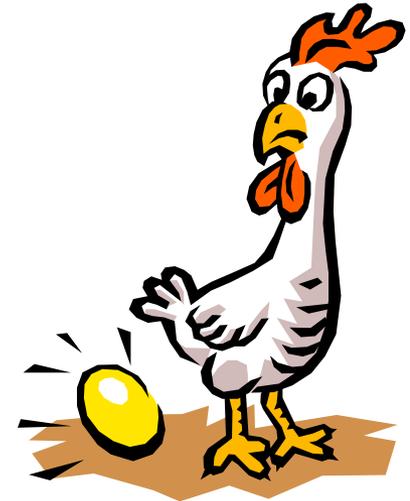


issues with alignment (cont'd)

- **evolutionary scenario** - not taken into account



- **statistical viewpoint** - hard to control biases created by alignment



literature overview

- **empirical work**
 - *issues with multiple alignment*
Loytynoja & Goldman, Science (2008); Wong et al, Science (2008)
 - *alignment-free methods*
E.g. Hohl & Ragan, Syst Biol (2007) and refs therein
 - *joint estimation of alignment and phylogeny*
Suchard & Redelings, Bioinformatics (2006); Liu et al, Science (2009); etc
- **theoretical work**
 - *word statistics*
E.g. Reinert et al., J Comput Biol (2000) and refs therein
 - *consistent estimation under TKF (?)*
Thatte, Math Biosci (2006)
 - *sequence-length requirements*
Erdos et al, Rand Struct Algor (1999); etc

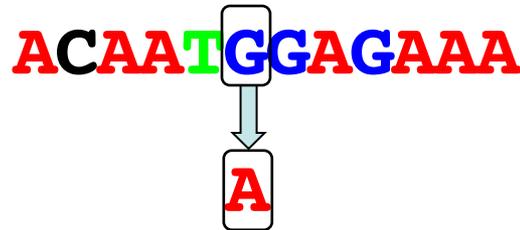
alignment-free reconstruction

- **theoretical result** [Daskalakis-R., Annals of Applied Probability, 2013]
 - we develop a **computationally efficient, statistically consistent** approach to **reconstruct** a species tree under the TKF indel process
 - does not require alignment
 - also give sequence-length requirements similar to substitution-only case
 - based on a probabilistic analysis of the indel process



indel process (a la TKF)

- **mutations** - rate matrix Q (per site; independently)



- **deletions** - rate μ (per site; independently)



- **insertions** - rate λ (per site; independently); insertion state is uniform



to simplify
 $\mu = \lambda$

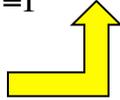
distance matrix: substitution-only case

- **data** - n aligned sequences

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

- $p'(a,b)$: **proportion** of sites that disagree between sequences a and b
 - example: $p'(\text{Homo sapiens}, \text{Pan}) = 0.2$
- **CFN formula** - map $\{A,G\}$ to +1 and $\{C,T\}$ to -1 and let $p'(a,b)$ be the corresponding proportion of disagreements

$$D'(a,b) = -\frac{1}{2} \log(1 - 2p'(a,b)) = -\frac{1}{2} \log\left(\frac{1}{k} \sum_{i=1}^k s_a^i s_b^i\right)$$

Expectation = $\exp(-2\mu t)$


- **goal** - find a distance for TKF model (that avoids the theoretical problems of edit distance and k-mer count distances)

no need for alignment: displacements are concentrated

- **single channel** - consider a path of length t

- each site survives with probability

$$\exp(-\mu t)$$

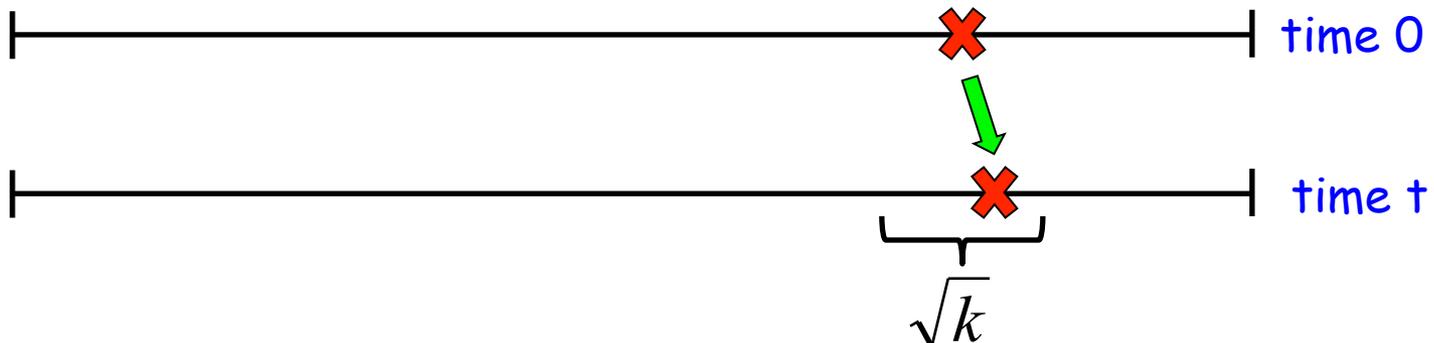
- so number of surviving sites is

$$k \exp(-\mu t) \pm \sqrt{k}$$

- similar argument for insertions implies total length is

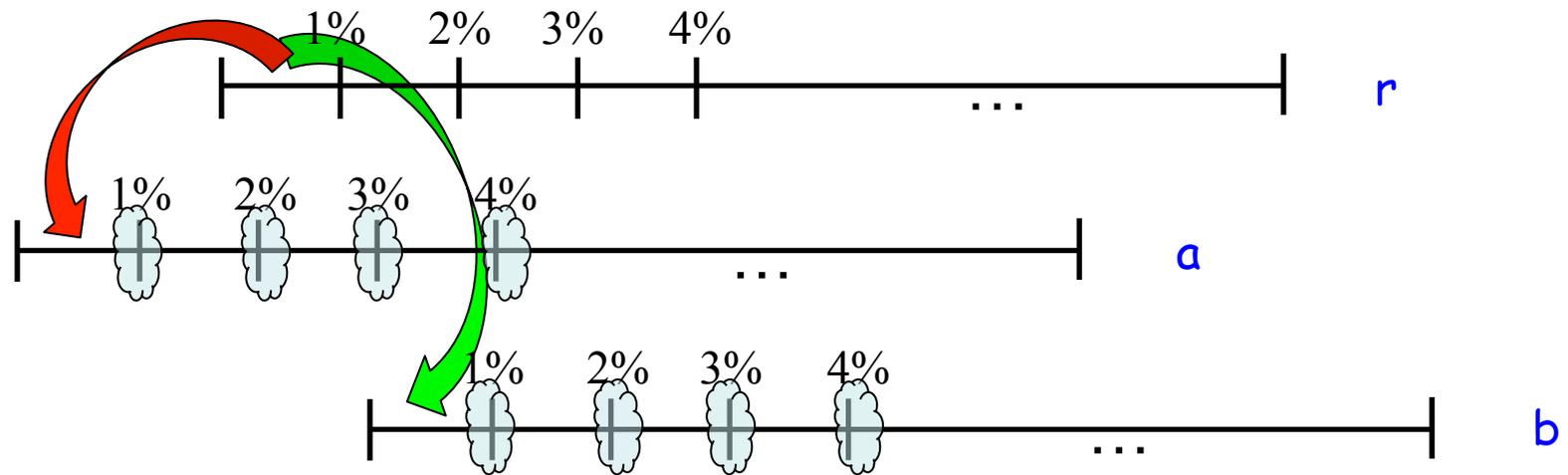
$$k \pm \sqrt{k}$$

- ALSO applies to **site locations**

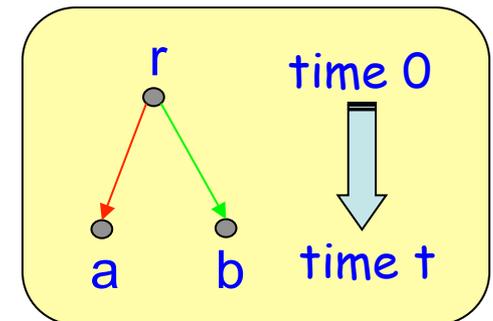


alignment-free comparison

- **looking from a distance** - divide the sequences into blocks



- reconstructed blocks may be off but only by a negligible fraction



block-wise statistics

- **single block** - consider a block of length K

- we use the agglomerated statistic

$$R_a^x = \sum_{i \in x} S_a^i$$

- divide into contributions from jointly surviving sites and inserted sites

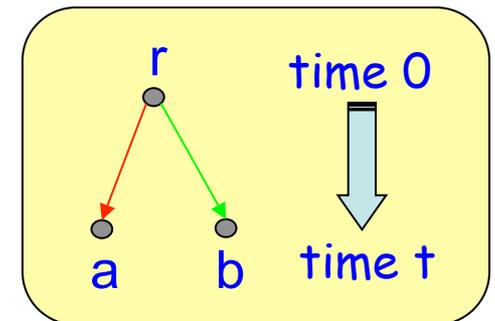
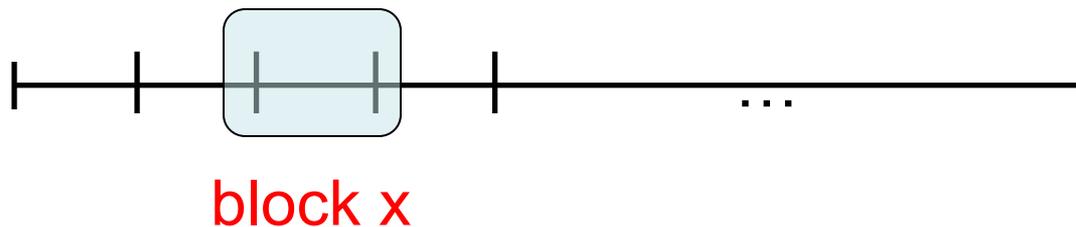
$$R_a^x R_b^x = \left(JS_a^x + I_a^x \right) \left(JS_b^x + I_b^x \right)$$

- jointly surviving sites contribute

$$K \exp(-2\mu t - 2qt) \pm K$$

- inserted sites contribute

$$0 \pm K$$



new distance: averaging over blocks

- **L blocks** are roughly independent

$$\frac{1}{L} \sum_x R_a^x R_b^x = \frac{1}{L} \sum_x \left(JS_a^x + I_a^x \right) \left(JS_b^x + I_b^x \right)$$

- **insertions** contribute on average 0

- **jointly surviving sites** contribute on average

$$K \exp(-2\mu t - 2qt)$$

- variance is roughly

$$K^2/L \ll K^2 \quad \text{if } L = \omega(1)$$

- so a **consistent time estimator** (at least proportional to time) is

$$-\frac{1}{2} \log \left(\frac{1}{L} \sum_x R_a^x R_b^x \right) \xrightarrow{L, k \rightarrow \infty} (\mu + q) t$$

summary

Theorem 1 (Consistency). *Assume that $0 < t_e, \eta_e < +\infty$, for all $e \in E$. Moreover, assume that the indel rates satisfy $\lambda_e < \mu_e$ for all $e \in E$. Under these assumptions, there exists an algorithm solving the phylogenetic reconstruction problem (that is, returning the correct tree) with probability of failure approaching 0 as the sequence length at the root of the tree goes to $+\infty$.*

Theorem 2 (Main Result: Two-State Ultrametric Case). *Assume there exist constants $0 < f, g < +\infty$, independent of n , such that all branch lengths t_e , $e \in E$, satisfy $f < t_e < g$. Moreover, assume that $\eta_e = \eta$, for all $e \in E$, where η is bounded between two constants $\underline{\eta} > 0$ and $\bar{\eta} < +\infty$ independent of n , and that the indel rates satisfy $\lambda_e = \lambda$, $\mu_e = \mu$, for all $e \in E$, and $\lambda < \mu = O(1/\log n)$. Under the assumptions above, there exists a polynomial-time algorithm solving the phylogenetic reconstruction problem (that is, returning the correct tree) with probability of failure $O\left(n^{-\beta'}\right)$, if the root sequence has length $k_r = \text{poly}_{\beta'}(n)$.*

thank
you