Mind the gap: Modelling evolution with indels

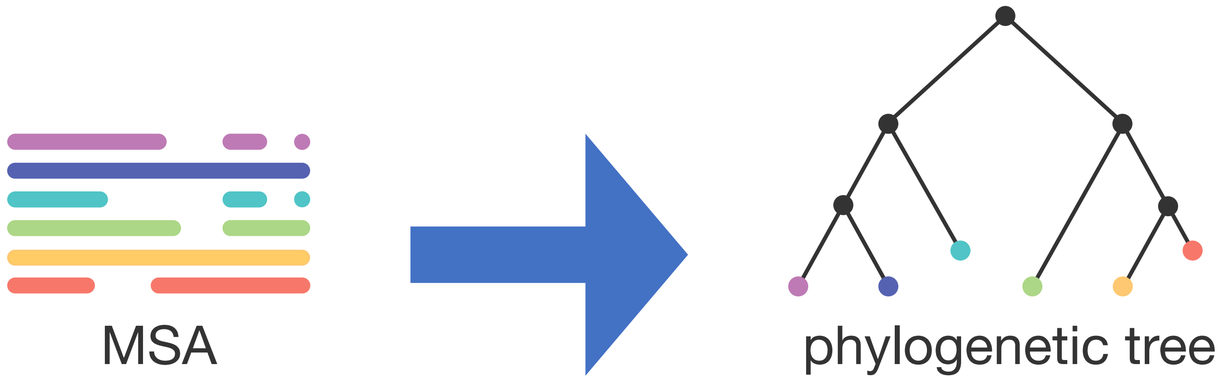
TAMING THE BEAST 2023

Jūlija Pečerska

17.08.23

PHYLOGENETIC PIPELINES

EXPECTATION

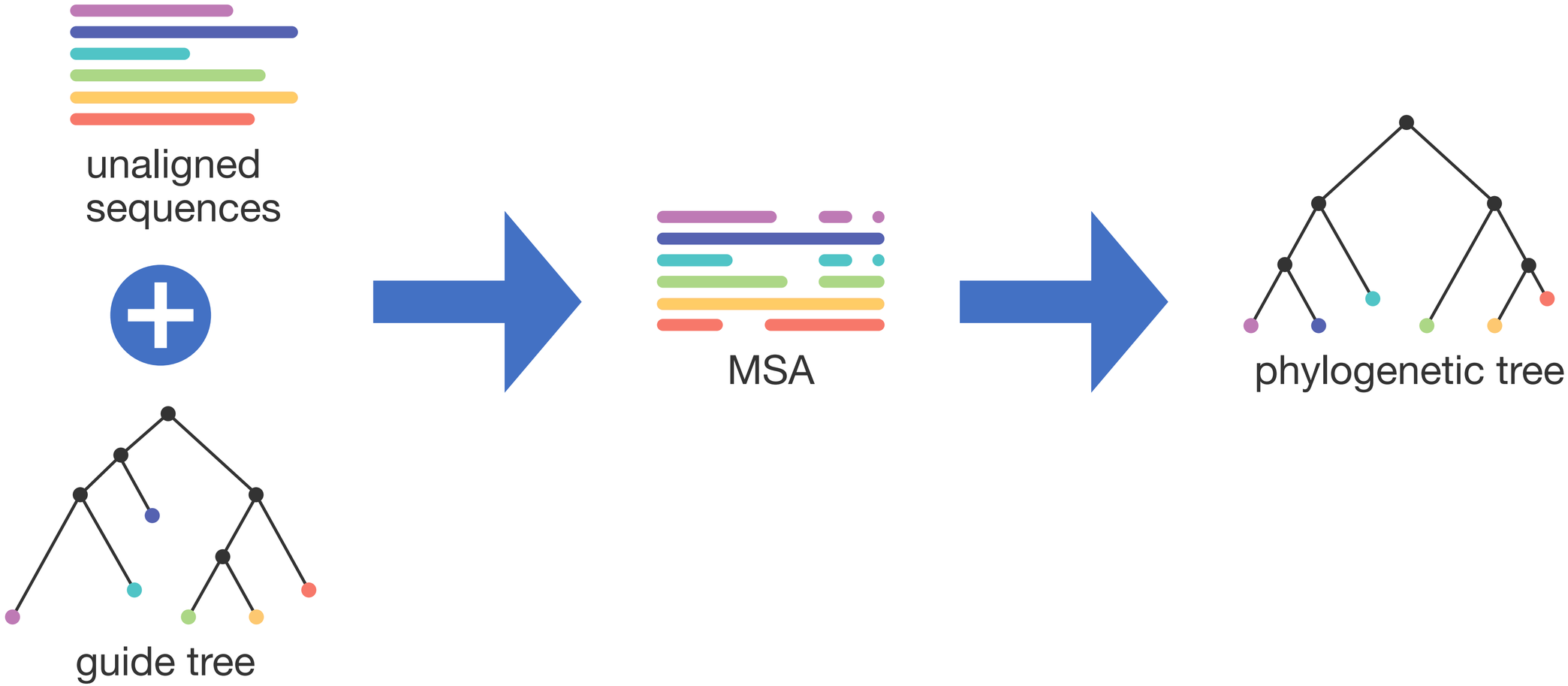


PHYLOGENETIC PIPELINES

REALITY

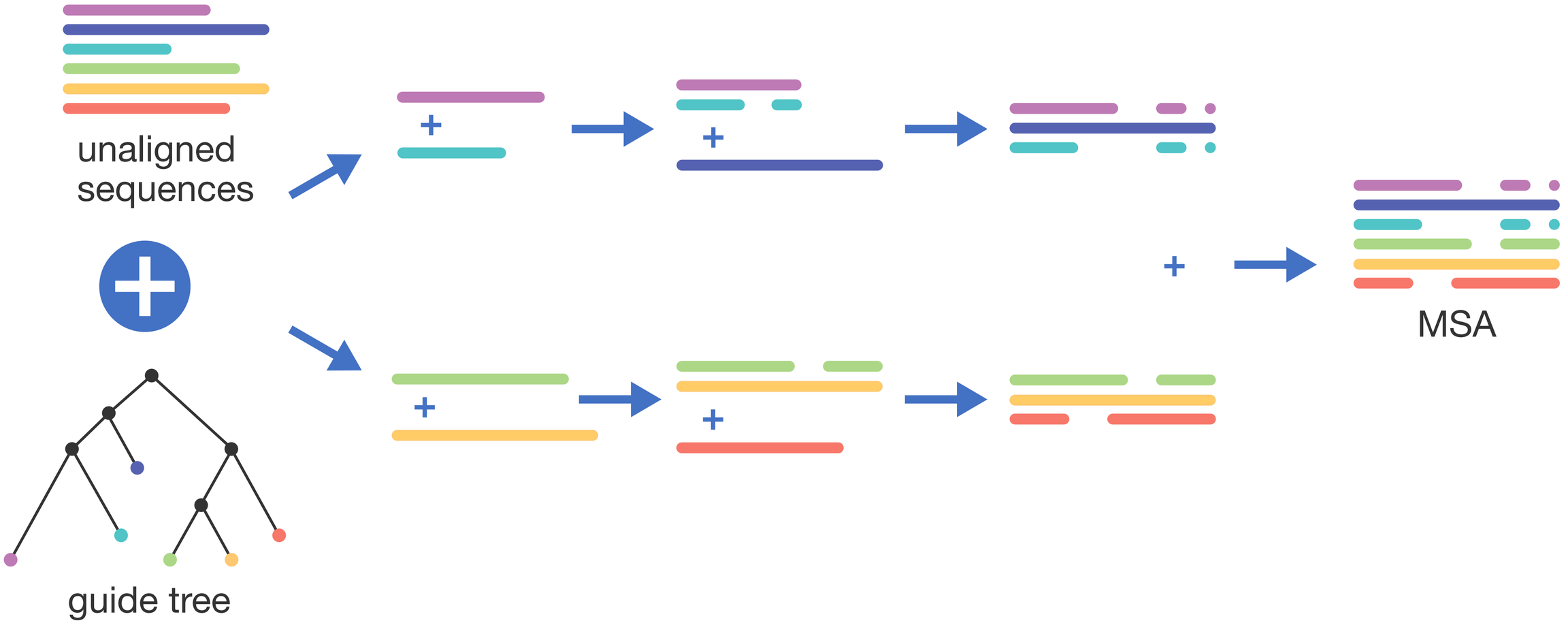
PHYLOGENETIC PIPELINES

REALITY



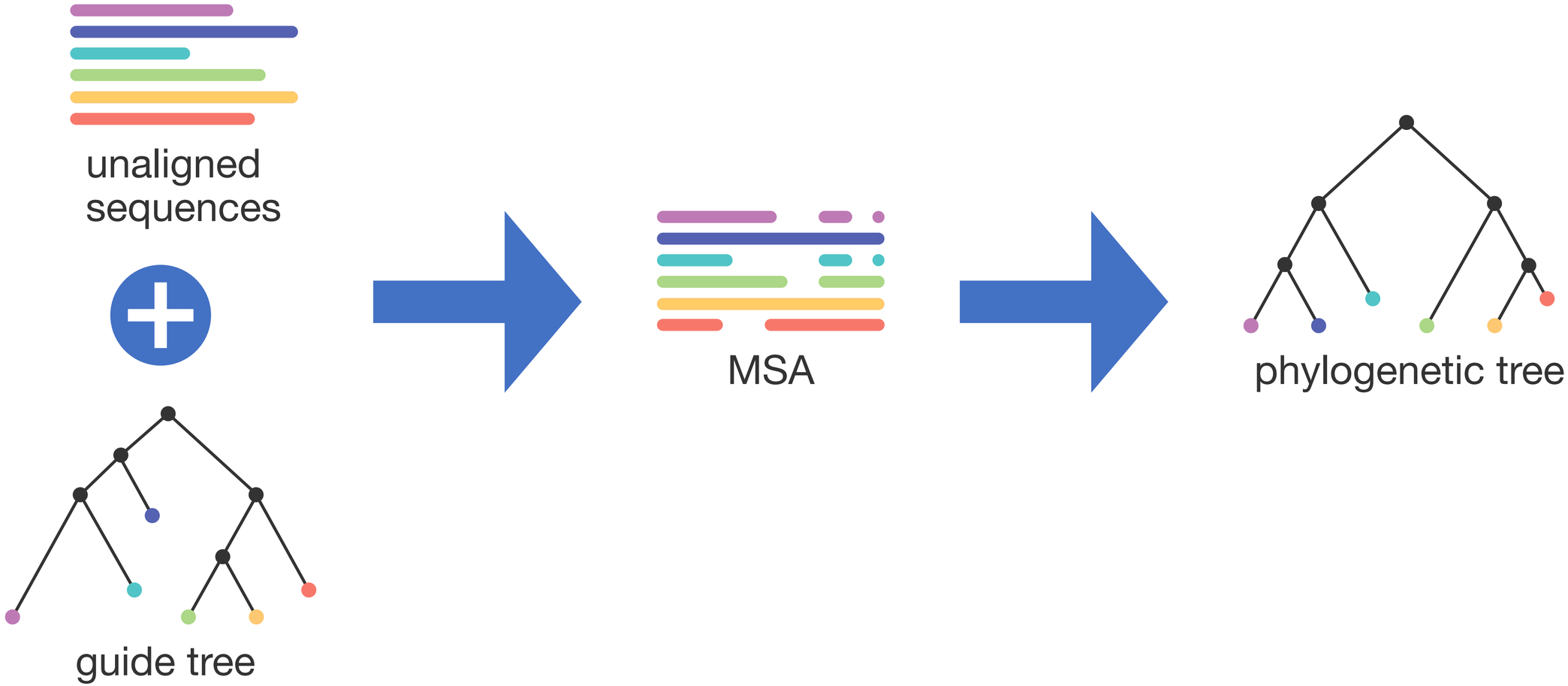
PHYLOGENETIC PIPELINES

REALITY



PHYLOGENETIC PIPELINES

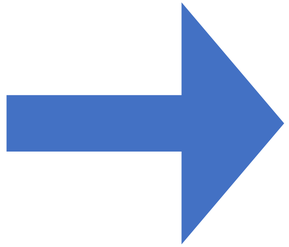
REALITY

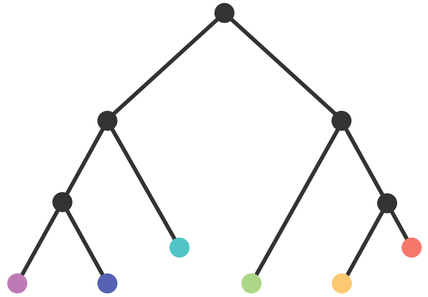


PHYLOGENETIC PIPELINES

REALITY

MSA





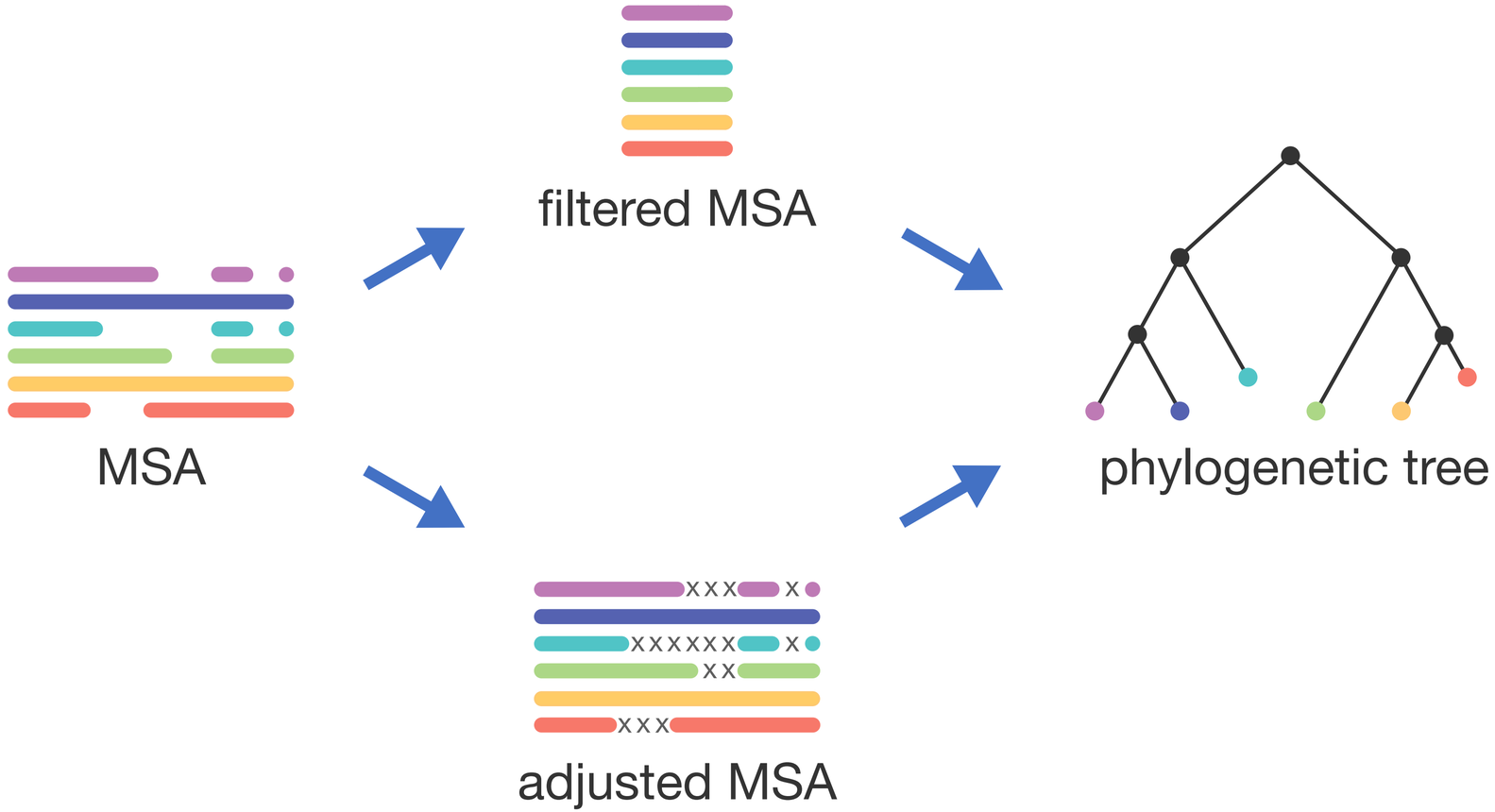
phylogenetic tree

PHYLOGENETIC PIPELINES

REALITY

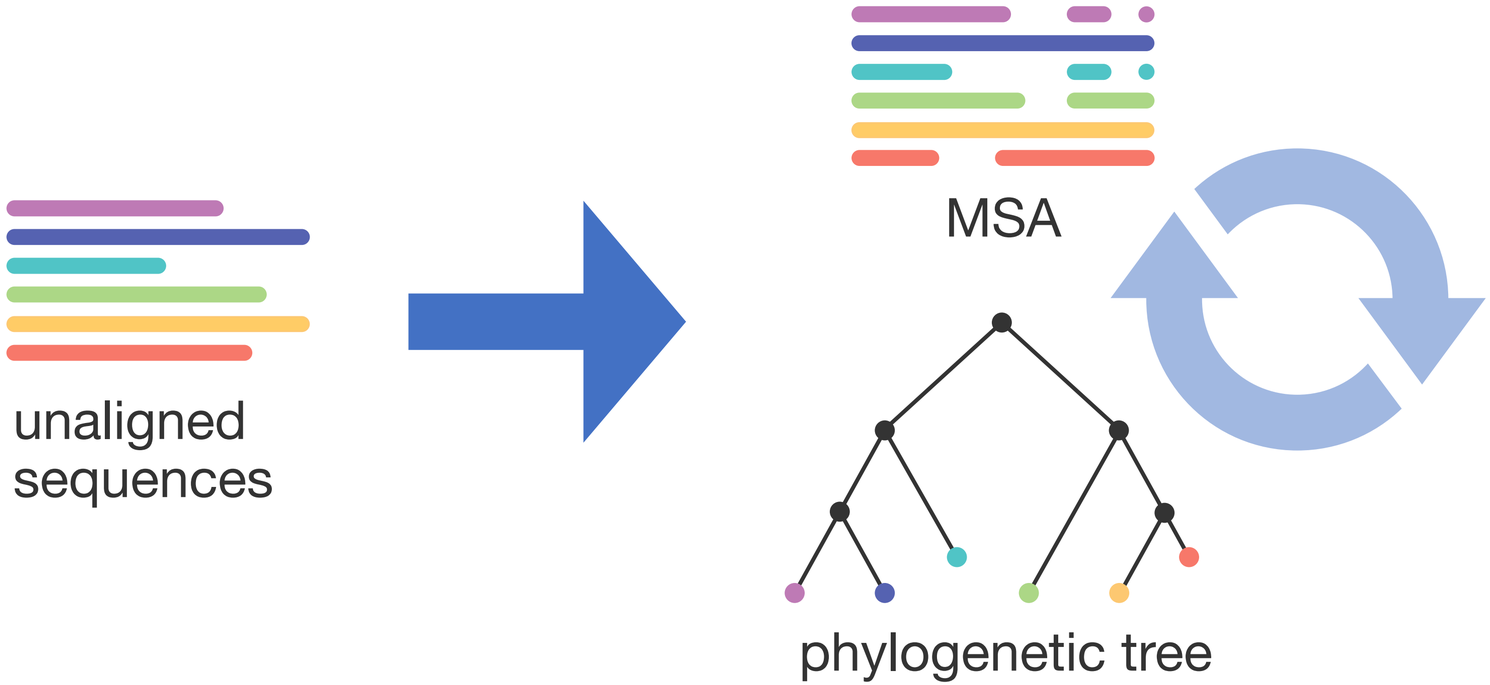
PHYLOGENETIC PIPELINES

REALITY



PHYLOGENETIC PIPELINES

EXPECTATION 2.0



REMINDER: SUBSTITUTION MODELS

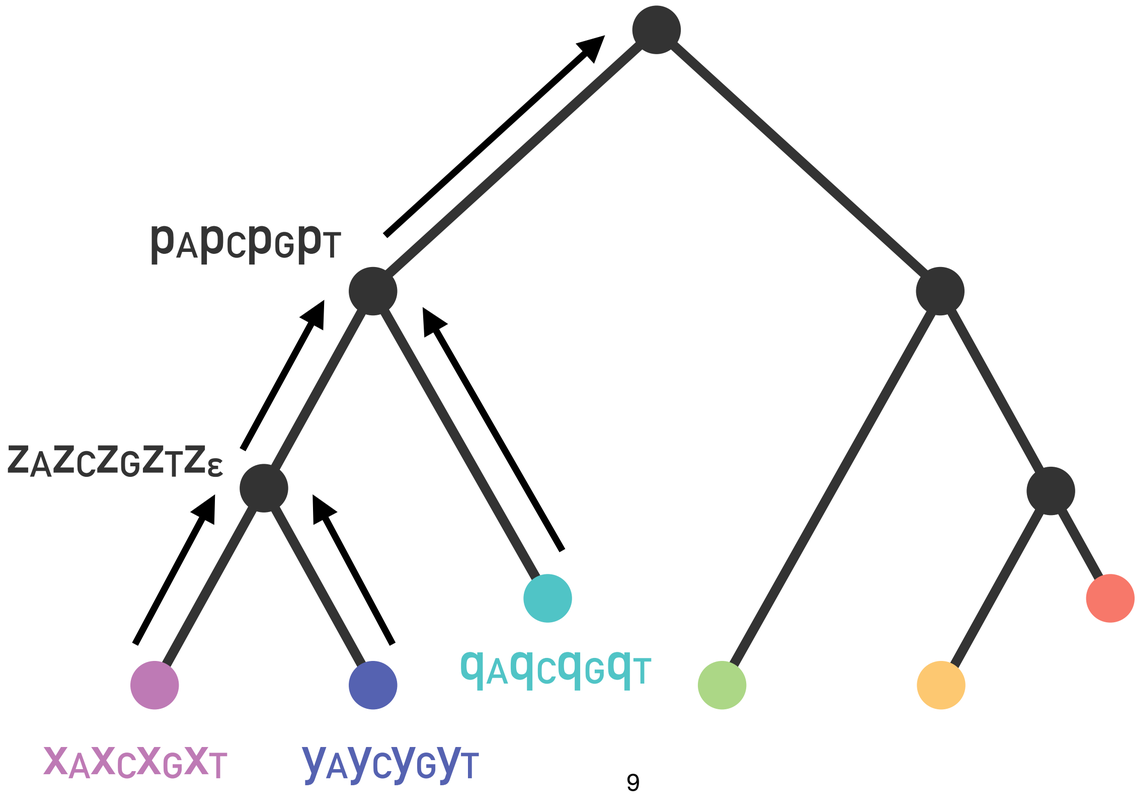
CONTINUOUS TIME MARKOV CHAINS

- Q — substitution rate matrix
- $\mathrm{P}(t)$ — transition probability matrix
- $\mathrm{P}(t) = e^{t\mathrm{Q}}$
  - $\mathrm{P}(t+s) = e^{(t+s)\mathrm{Q}} = e^{t\mathrm{Q}}e^{s\mathrm{Q}} = \mathrm{P}(t)\mathrm{P}(s)$

	T	C	A	G
T	…	$a\pi_{\mathrm{C}}$	$b\pi_{\mathrm{A}}$	$c\pi_{\mathrm{G}}$
C	$a\pi_{\mathrm{T}}$	…	$d\pi_{\mathrm{A}}$	$e\pi_{\mathrm{G}}$
A	$b\pi_{\mathrm{T}}$	$d\pi_{\mathrm{C}}$	…	$f\pi_{\mathrm{G}}$
G	$c\pi_{\mathrm{T}}$	$e\pi_{\mathrm{C}}$	$f\pi_{\mathrm{A}}$	…

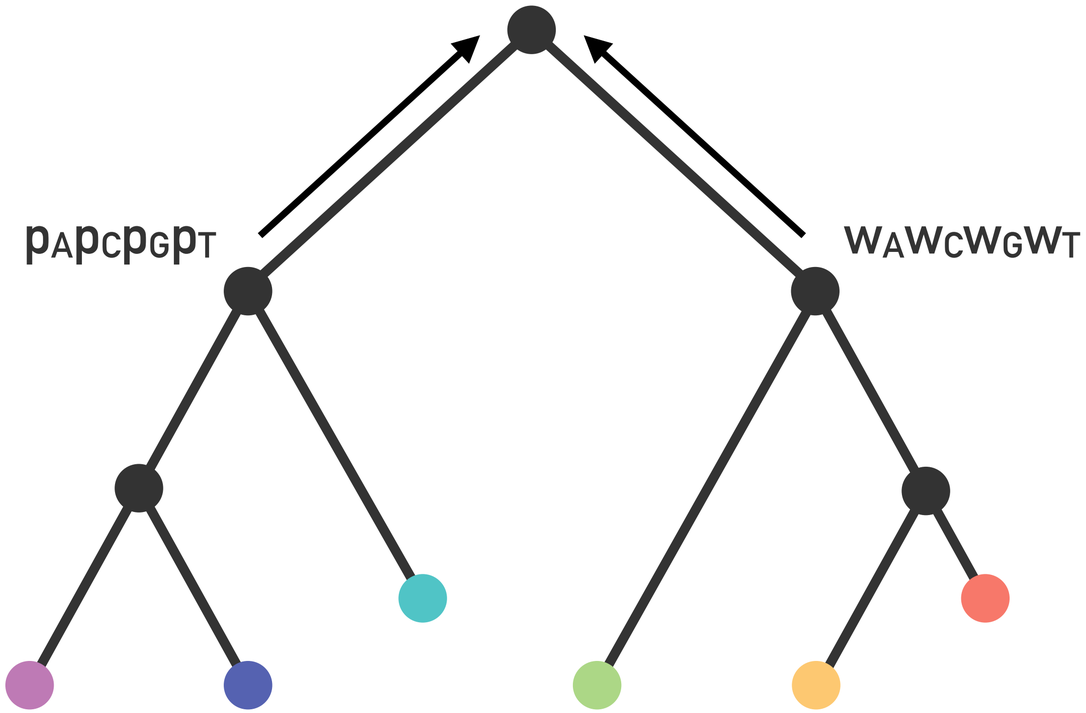
REMINDER: PHYLOGENETIC LIKELIHOOD

USING FELSENSTEIN'S PRUNING ALGORITHM



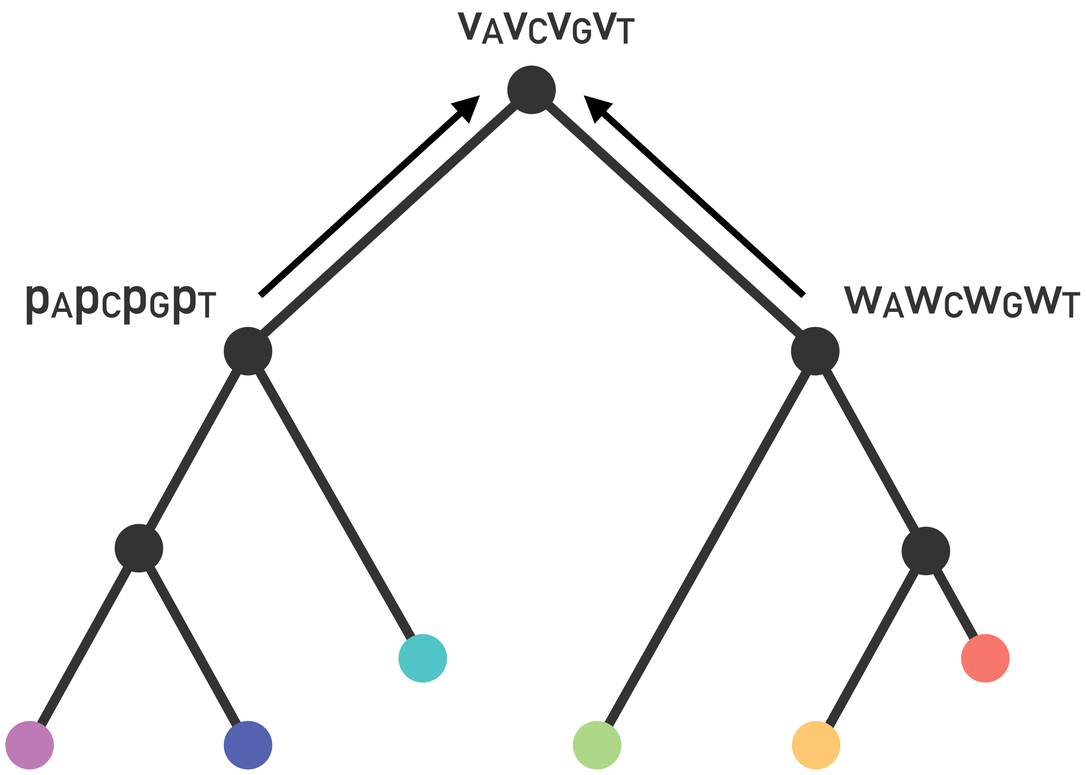
REMINDER: PHYLOGENETIC LIKELIHOOD

USING FELSENSTEIN'S PRUNING ALGORITHM



REMINDER: PHYLOGENETIC LIKELIHOOD

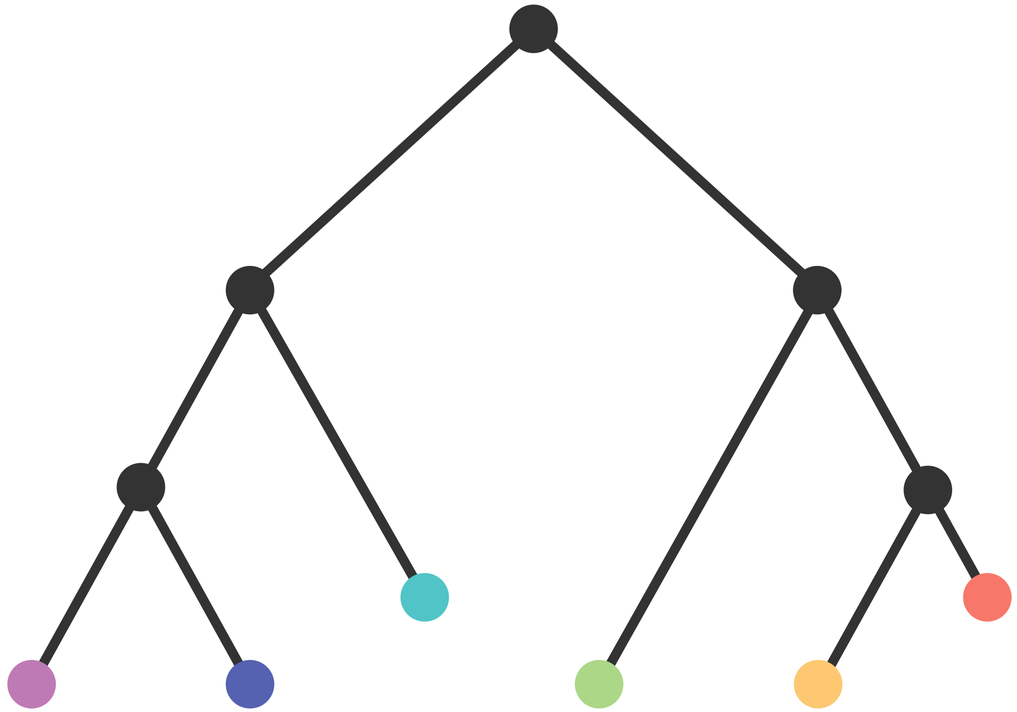
USING FELSENSTEIN'S PRUNING ALGORITHM



REMINDER: PHYLOGENETIC LIKELIHOOD

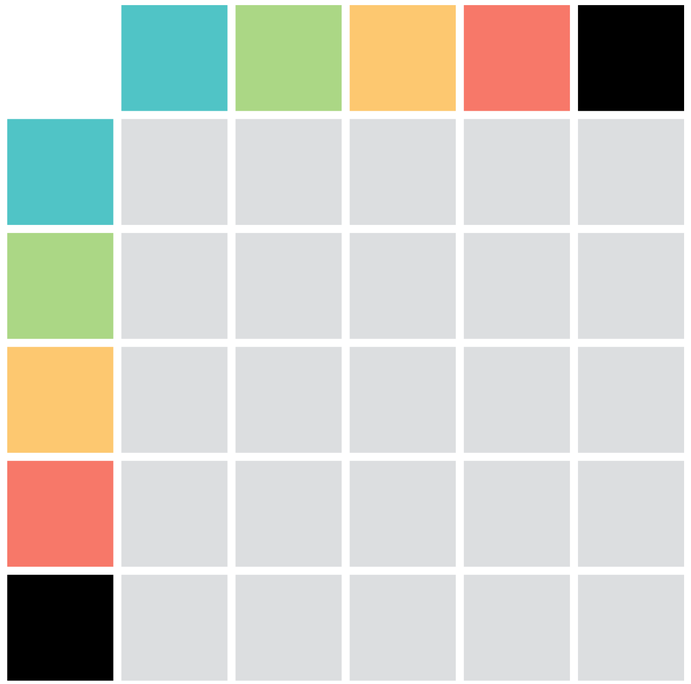
USING FELSENSTEIN'S PRUNING ALGORITHM

Likelihood = $f(v_A v_C v_G v_T)$



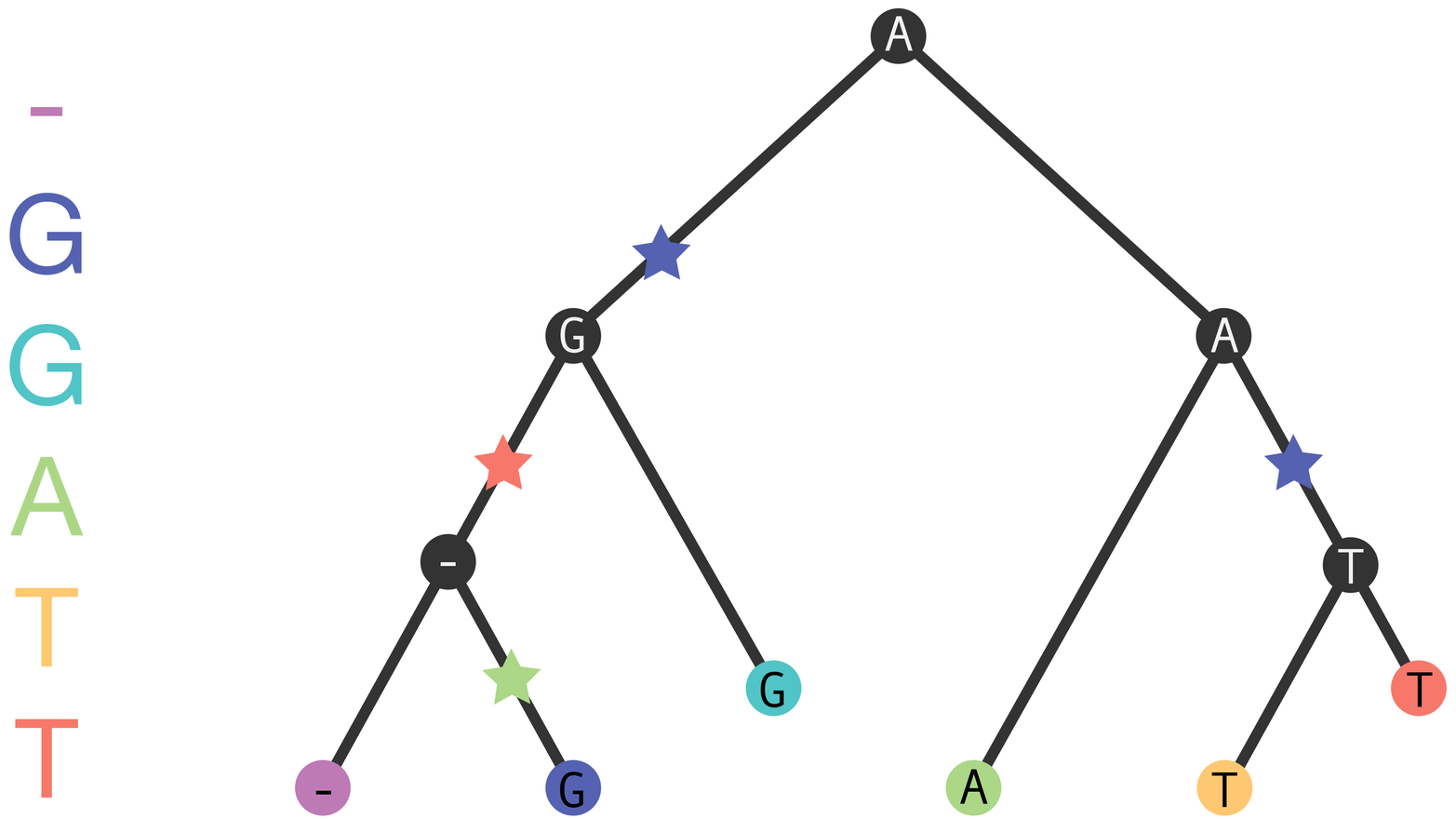
MODELLING INDELS

WHY NOT WITH MARKOV CHAINS?



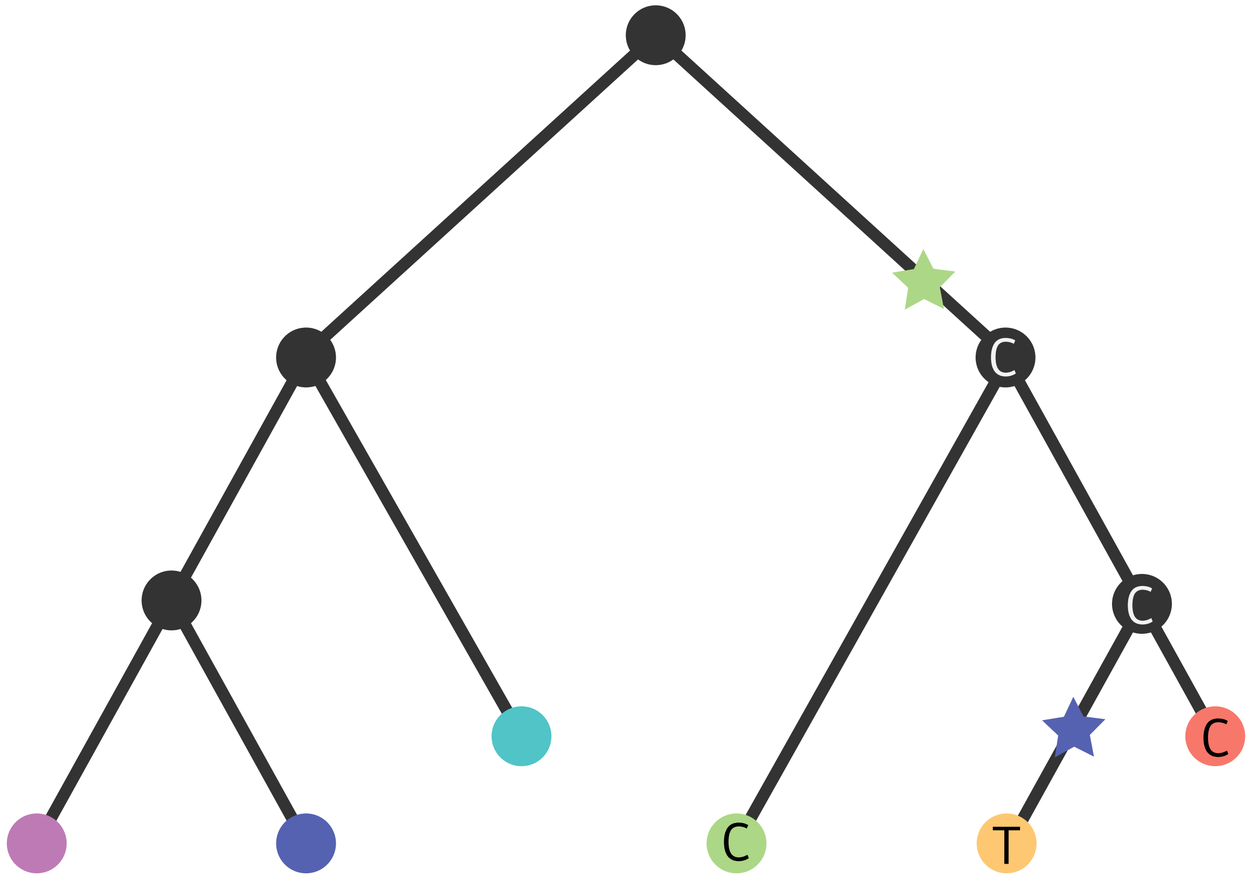
MODELLING INDELS

WHY NOT WITH MARKOV CHAINS?

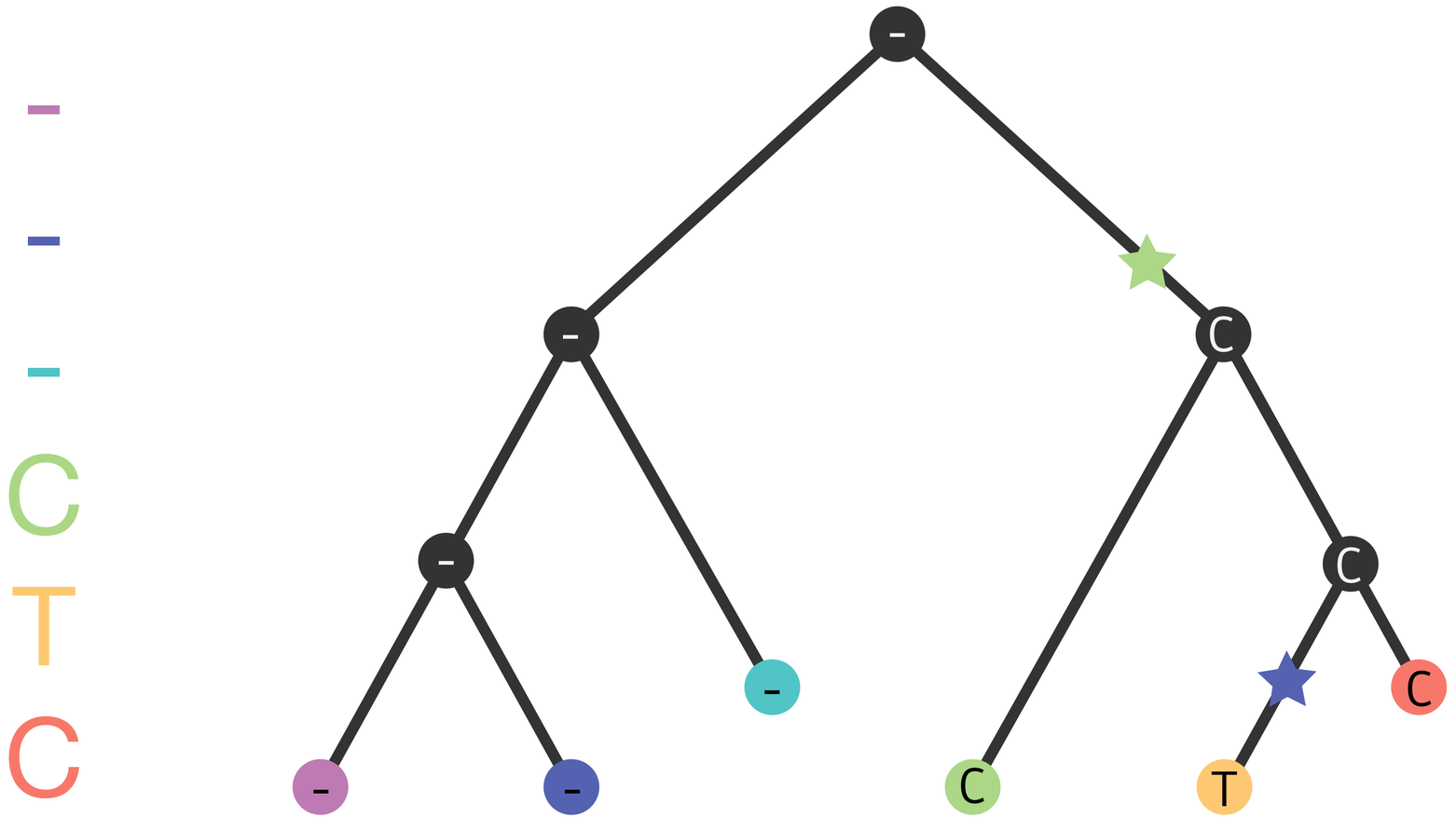


ANOTHER FUNKY INDEL SCENARIO

C
T
C

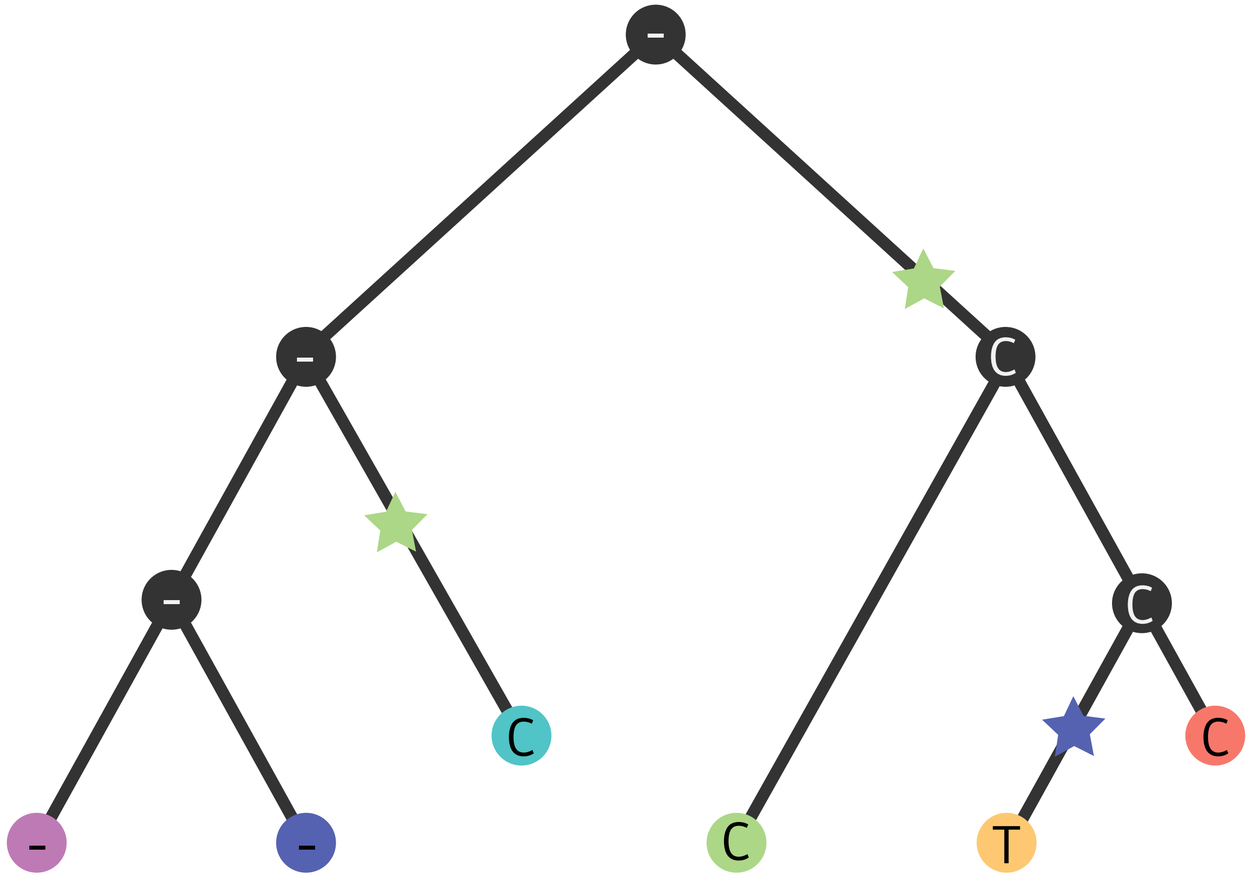


ANOTHER FUNKY INDEL SCENARIO



ANOTHER FUNKY INDEL SCENARIO

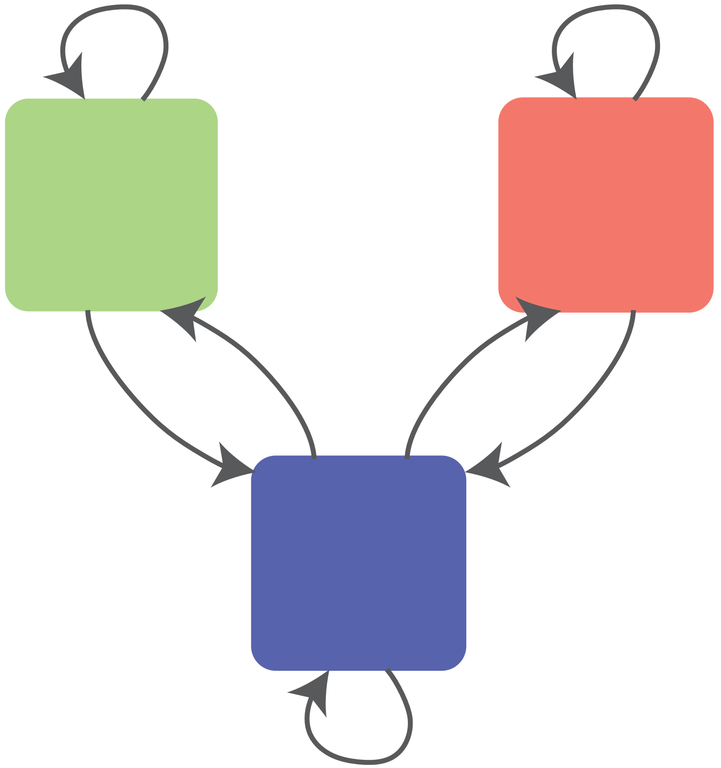
- - C C T C



MODELLING INDELS

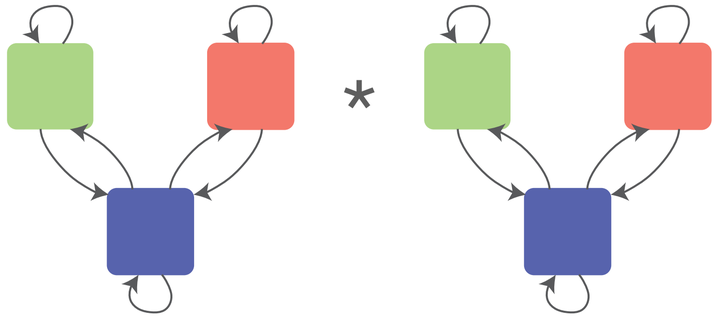
USING HMMS

- Process (e.g. TKF91/92) described with an HMM



MODELLING INDELS

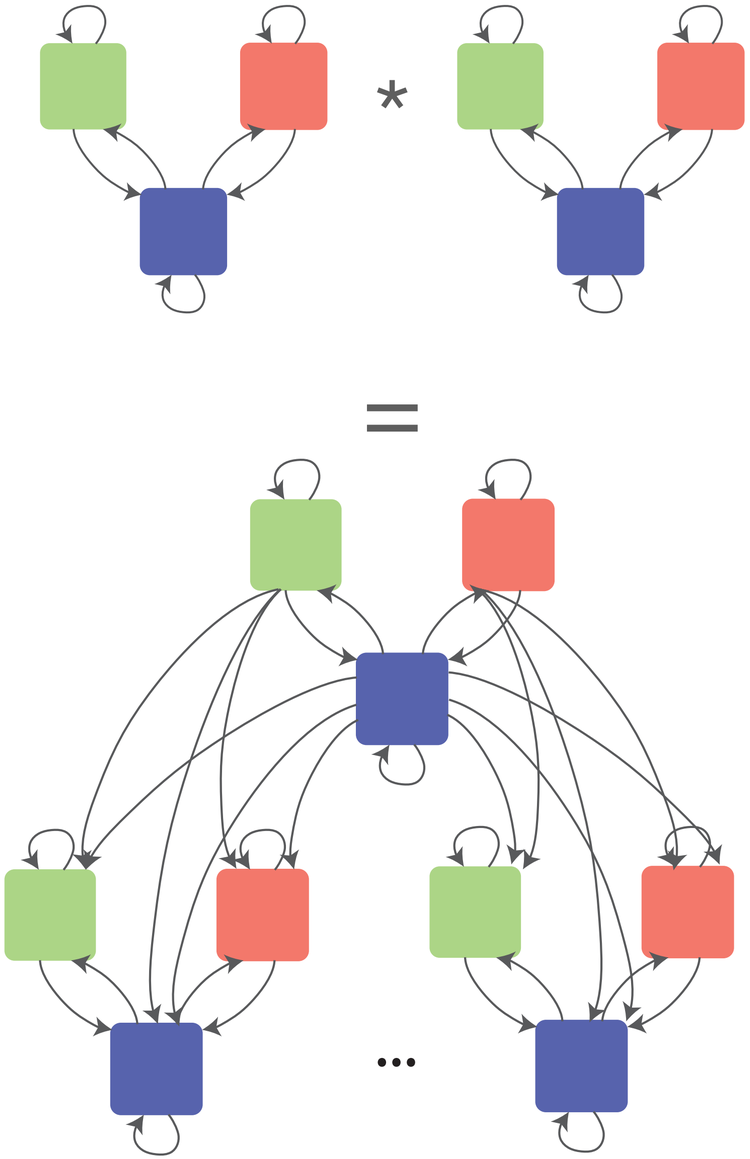
USING HMMS



- Process (e.g. TKF91/92) described with an HMM
- Differentiating along a branch explodes the HMM

MODELLING INDELS

USING HMMS

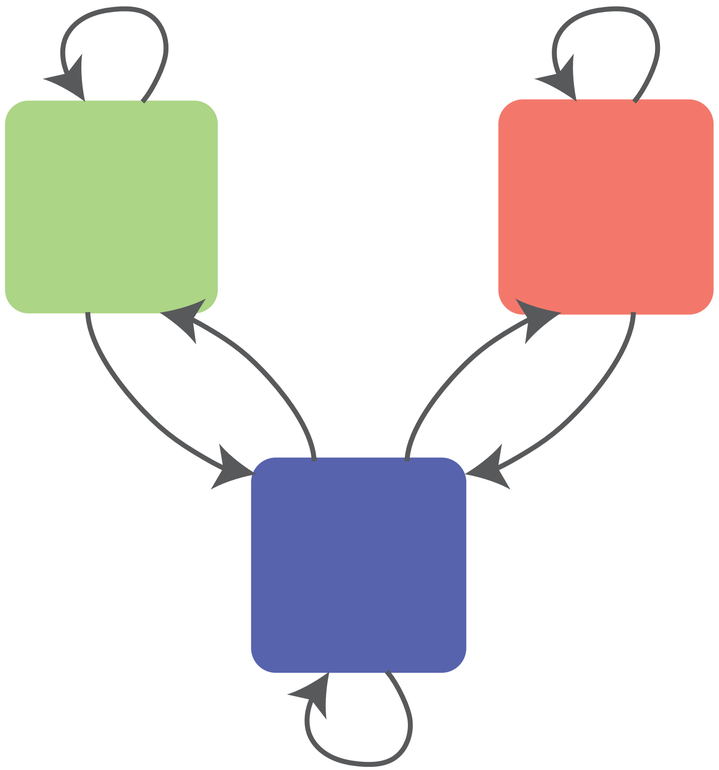


- Process (e.g. TKF91/92) described with an HMM
- Differentiating along a branch explodes the HMM
  - Finite branch length -> infinite states

MODELLING INDELS

USING HMMS

- Process (e.g. TKF91/92) described with an HMM
- Differentiating along a branch explodes the HMM
  - Finite branch length -> infinite states
- Can be simplified!
  - For local likelihood;
  - Global likelihood complexity is still exponential.



MODELLING INDELS...

... is very hard

MODELLING INDELS

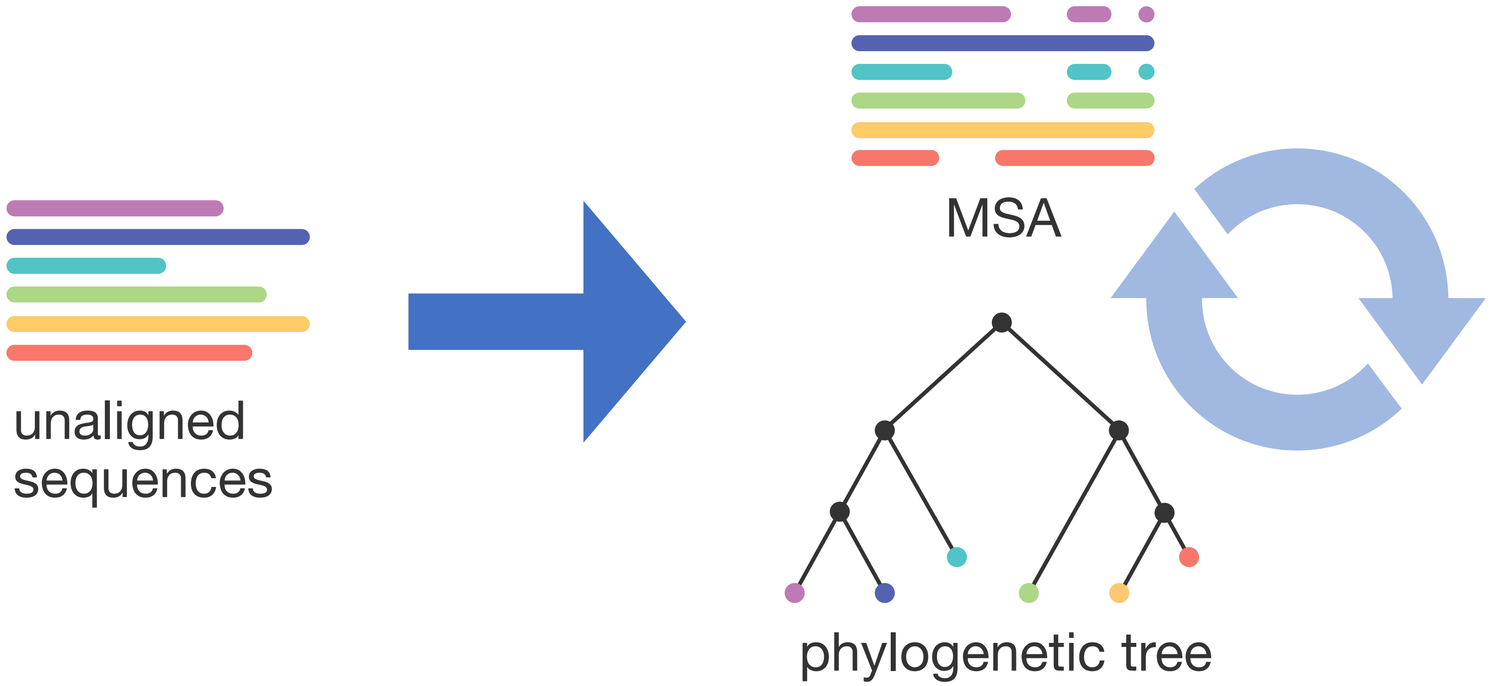
POISSON INDEL PROCESS (PIP)

PIP = Poisson insertion process on a tree
+ Substitution/deletion Markov process

- Properties:
  - Single character insertions/deletions
  - Time-reversible
  - Linear time likelihood

JOINT INFERENCE: THE DREAM

EXPECTATION 2.0



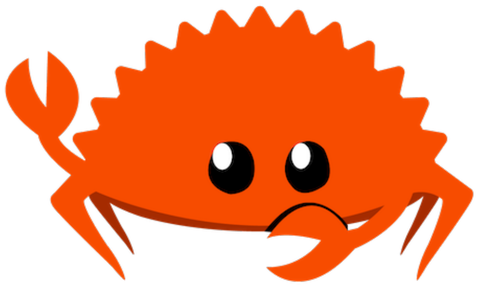
BAli---Phy

JOINT INFERENCE

REALITY

- JATI 1.0 necessary (but not sufficient) conditions:
  - Efficient
    - MSA adjustment in $O(NL^3)$
    - Tree search in $O(N^2L)$
  - Open-source
  - User-friendly

- Implementation:
  - Rust codebase (github.com/acg-team/JATI (private for now))
  - Release information (twitter.com/JulijaPecerska)



TAKEAWAYS

- Indel models are complicated;
- Know your data!!!
- Know your models and priors;
- Know your software and its assumptions.

ACKNOWLEDGEMENTS

- ACGTeam @ ZHAW
- Prof Maria Anisimova 
- Dr Manuel Gil 
- Clara Iglhaut 
- Gholamhossein Jowkar 

Swiss National Science Foundation