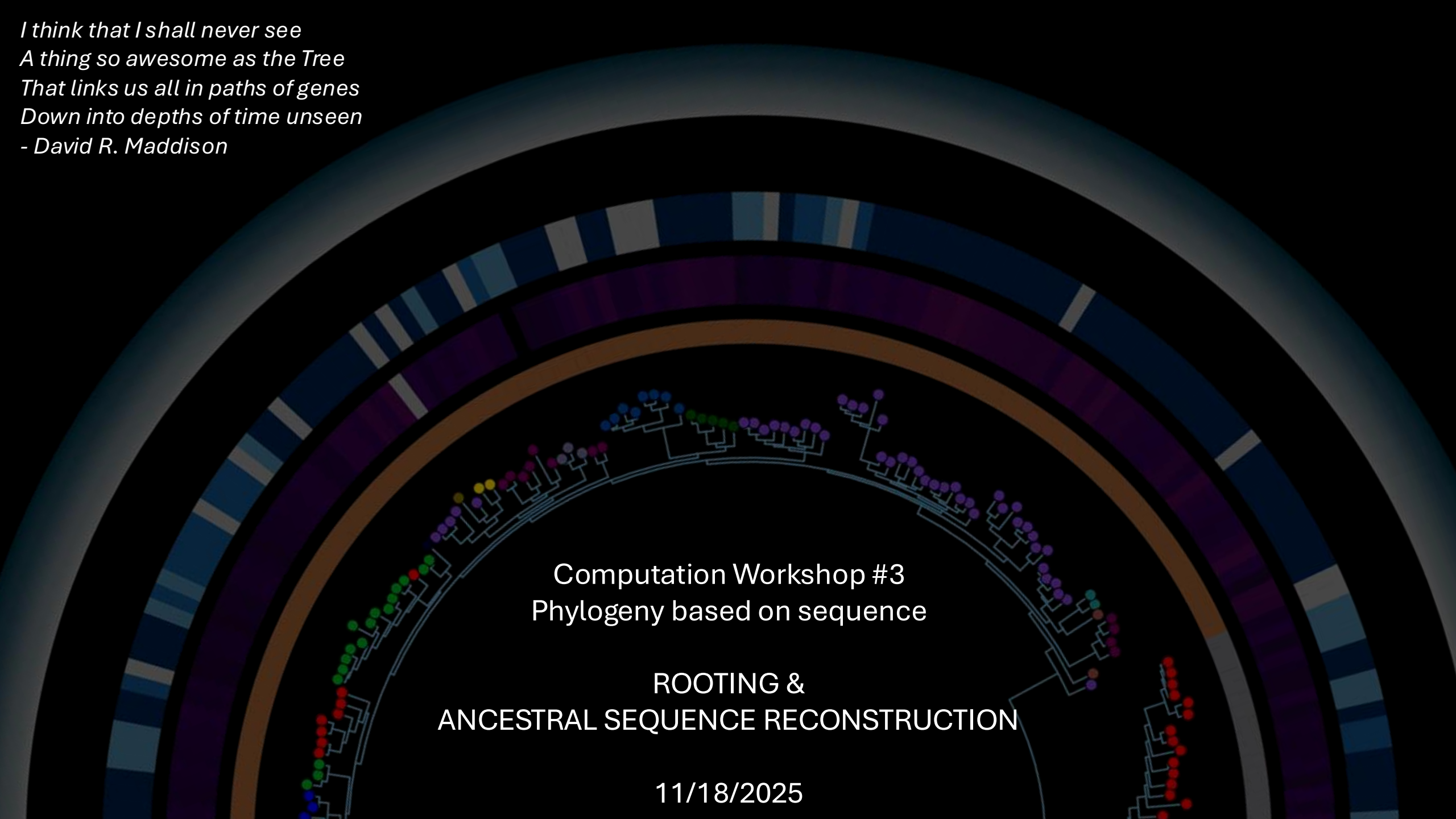
I think that I shall never see
A thing so awesome as the Tree
That links us all in paths of genes
Down into depths of time unseen
- David R. Maddison

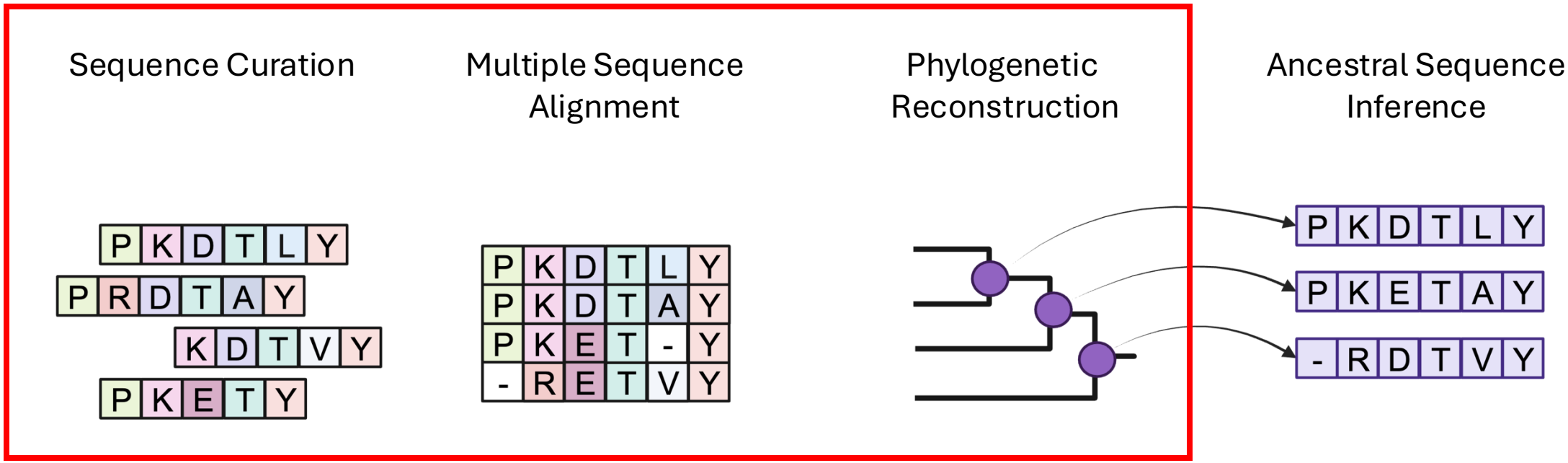


Computation Workshop #3
Phylogeny based on sequence

ROOTING &
ANCESTRAL SEQUENCE RECONSTRUCTION

11/18/2025

ASR Workflow



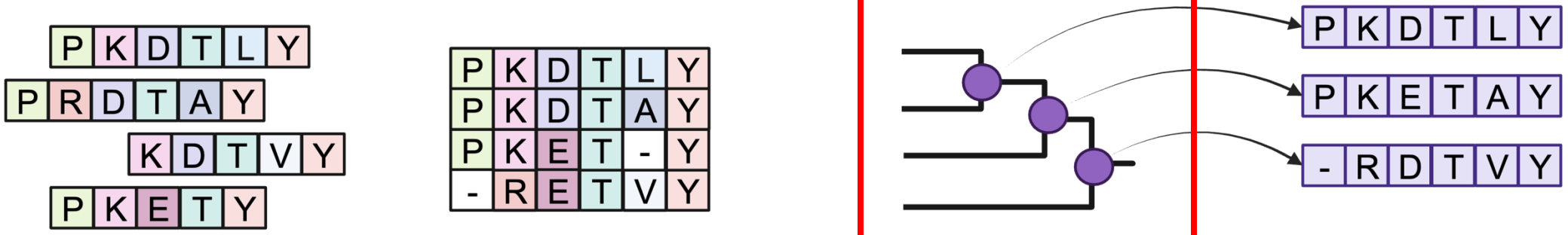
ASR Workflow

Sequence Curation

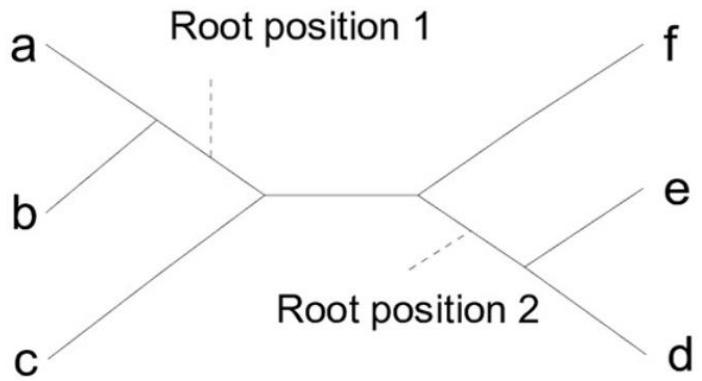
Multiple Sequence Alignment

Phylogenetic Reconstruction

Ancestral Sequence Inference

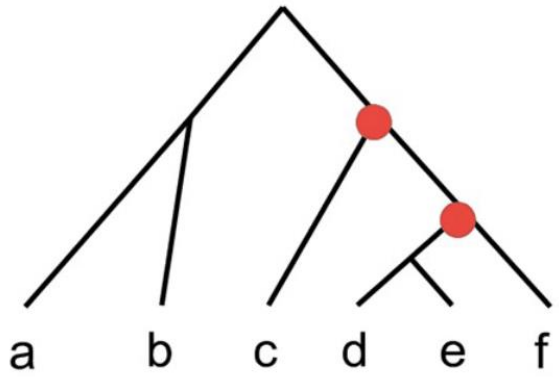


Unrooted vs Rooted Tree

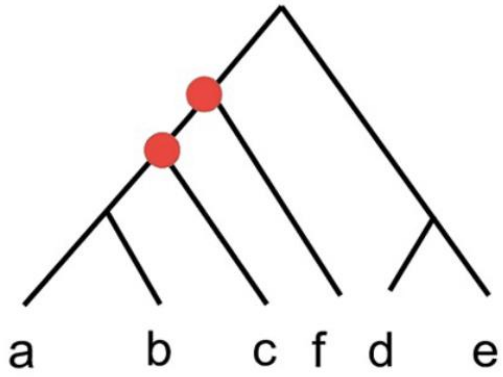


Unrooted Gene Tree

Rooting



Gene tree rooted at position 1



Gene tree rooted at position 2

No ancestor-descendant relationship

Ancestral relationships are determined

How to decide where to place the root? 🤷🏻‍♀️

Wade et al., *PLOS One*, 2020

Phylogenetic Tree Rooting Methods

Species Tree Rooting

Outgroup rooting

Phylogenetic reconciliation

Based on Branch Lengths

Midpoint rooting

Minimal variance

Minimal Ancestral Deviation

Using Molecular Clock

Molecular clock rooting

Bayesian molecular clock rooting

Non-reversible evolutionary models

Relaxed clock models

Phylogenetic Reconciliation for Gene Trees

Duplication-Transfer-Loss (DTL) reconciliation

ALE rooting

Wade et al., *PLOS One*, 2020

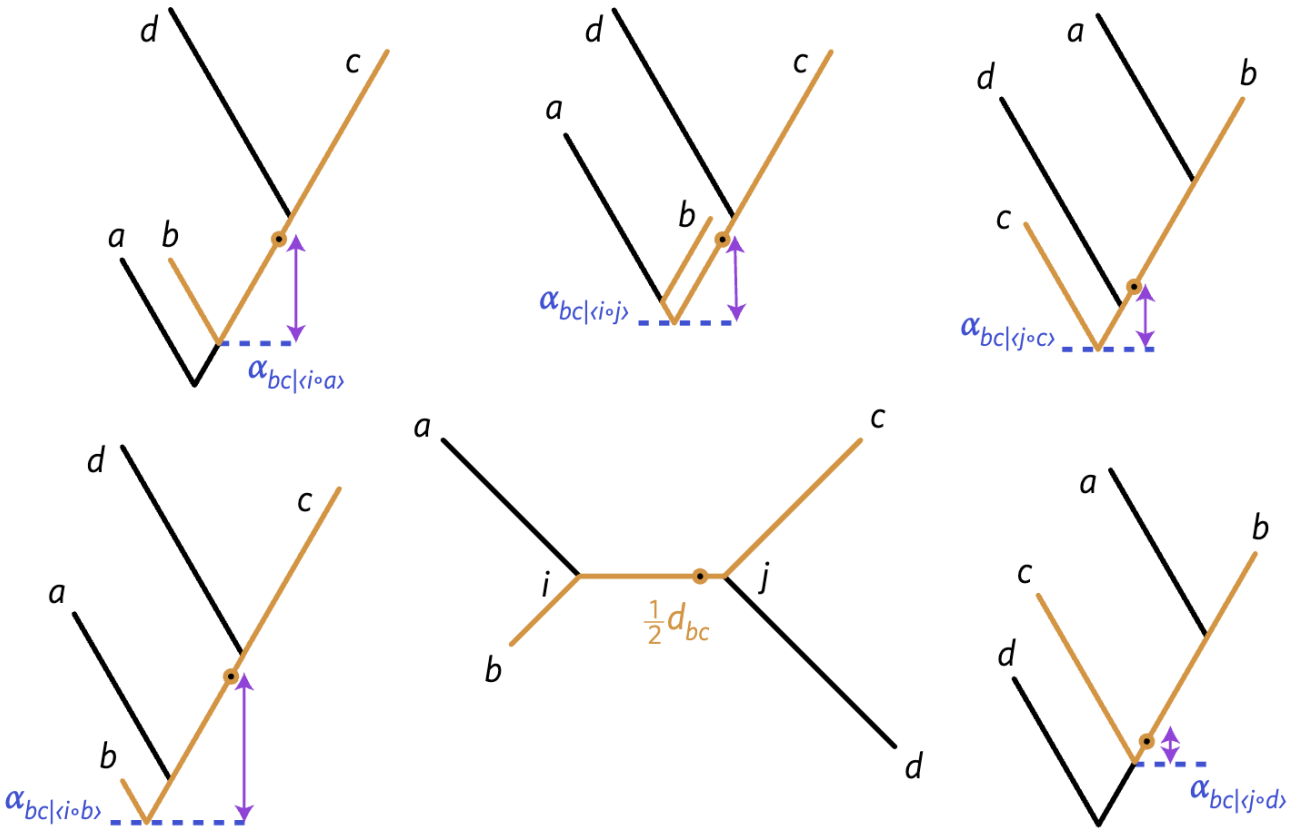
Rooting based on branch lengths

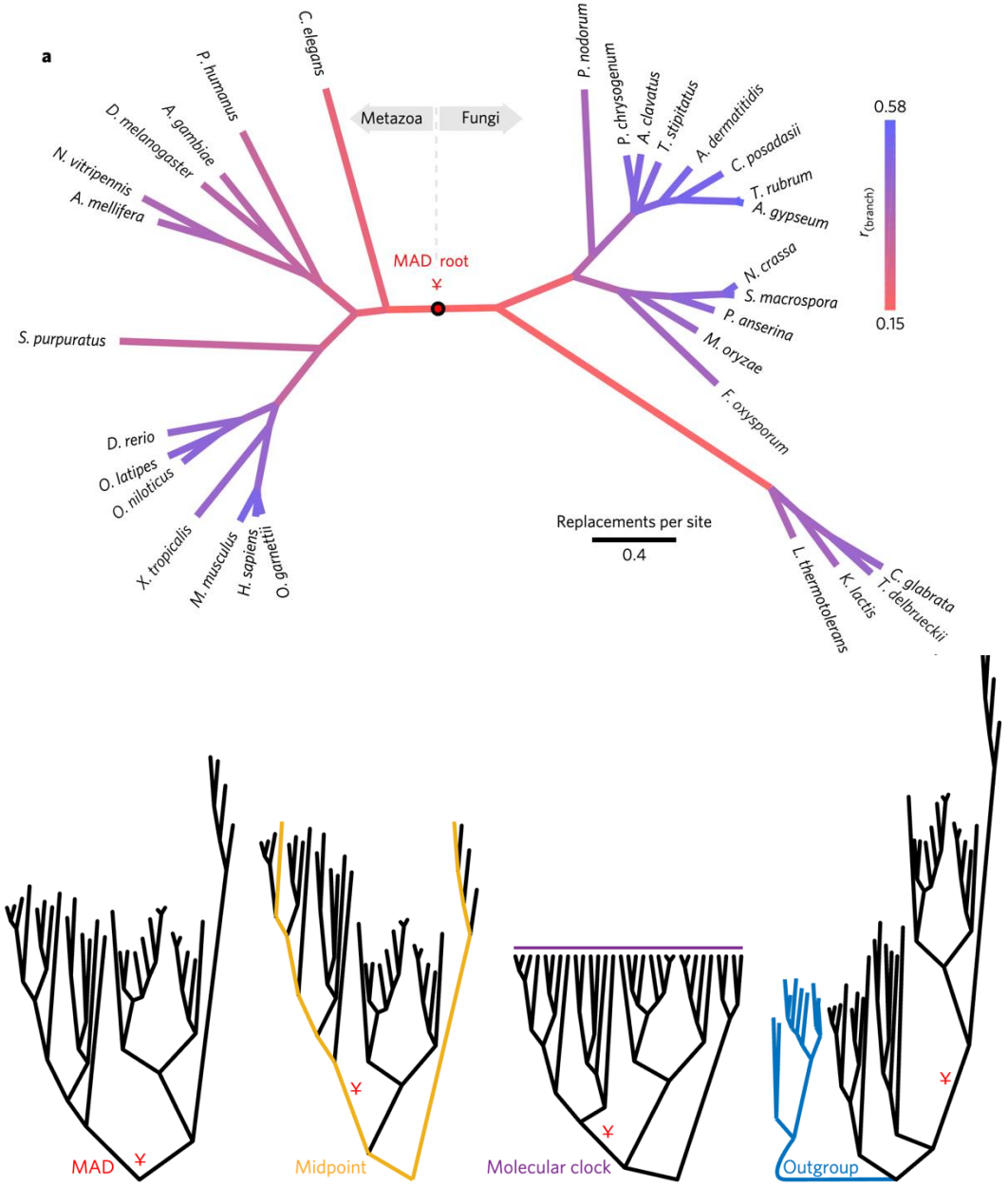
- **Midpoint**: unrooted tree is rooted at the mid-point of the longest path in the tree (Farris, *American Naturalist*, 1972)

- **Minimum Variance**: improve upon midpoint rooting, a root that minimizes root-to-leaf distance variance (Mai et al., *PLOS One*, 2017)

- **Minimal Ancestor Deviation (MAD)**: deviations of the midpoint criterion for all possible root positions (Tria et al., *Nat Eco Evo*, 2017)

How does MAD work?





Tria et al., *Nat Eco Evo*, 2017

Which one to use?

RESEARCH ARTICLE

Assessing the accuracy of phylogenetic rooting methods on prokaryotic gene families

Taylor Wade[1], L. Thiberio Rangel[2], Soumya Kundu[1¤], Gregory P. Fournier[2], Mukul S. Bansal[1,3]*

1 Department of Computer Science & Engineering, University of Connecticut, Storrs, CT, United States of America, **2** Department of Earth, Atmospheric & Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States of America, **3** Institute for Systems Genomics, University of Connecticut, Storrs, CT, United States of America

Abstract

Almost all standard phylogenetic methods for reconstructing gene trees result in *unrooted* trees; yet, many of the most useful applications of gene trees require that the gene trees be correctly rooted. As a result, several computational methods have been developed for inferring the root of unrooted gene trees. However, the accuracy of such methods has never been systematically evaluated on prokaryotic gene families, where horizontal gene transfer is often one of the dominant evolutionary events driving gene family evolution. In this work, we address this gap by conducting a thorough comparative evaluation of five different rooting methods using large collections of both simulated and empirical prokaryotic gene trees. Our simulation study is based on 6000 true and reconstructed gene trees on 100 species and characterizes the rooting accuracy of the four methods under 36 different evolutionary conditions and 3 levels of gene tree reconstruction error. The empirical study is based on a large, carefully designed data set of 3098 gene trees from 504 bacterial species (406 Alphaproteobacteria and 98 Cyanobacteria) and reveals insights that supplement those gleaned from the simulation study. Overall, this work provides several valuable insights into the accuracy of the considered methods that will help inform the choice of rooting methods to use when studying microbial gene family evolution. Among other findings, this study identifies parsimonious Duplication-Transfer-Loss (DTL) rooting and Minimal Ancestor Deviation (MAD) rooting as two of the most accurate gene tree rooting methods for prokaryotes and specifies the evolutionary conditions under which these methods are most accurate, demonstrates that DTL rooting is highly sensitive to high evolutionary rates and gene tree error, and that rooting methods based on branch-lengths are generally robust to gene tree reconstruction error.

Tutorial time

- Running MAFFT

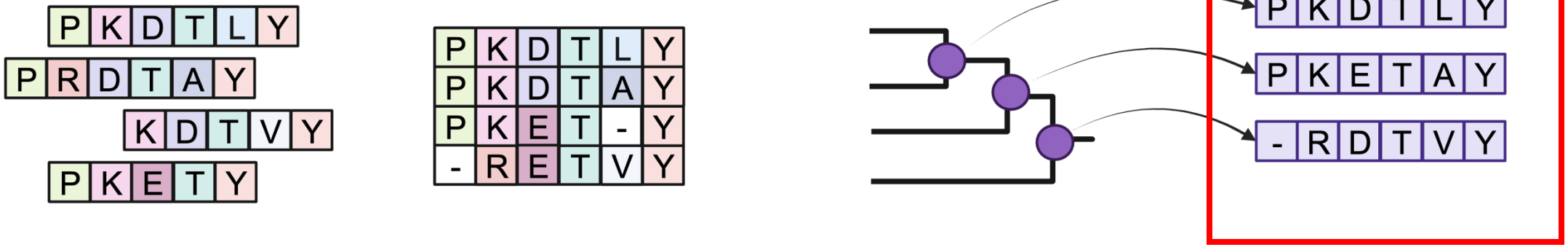
ASR Workflow

Sequence Curation

Multiple Sequence Alignment

Phylogenetic Reconstruction

Ancestral Sequence Inference

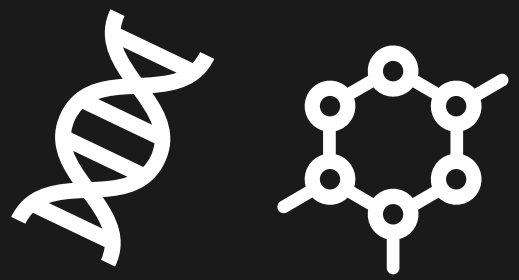


First time proposing the idea of reconstructing the ancestral genes/proteins

Two records of life:

1) Fossil Records

2) Molecular Records



Chemical Paleogenetics

Molecular "Restoration Studies" of Extinct Forms of Life

LINUS PAULING and EMILE ZUCKERKANDL*

*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California, USA***

Attention is attracted to the possibility of reconstructing the amino-acid sequence of ancestral polypeptide chains by virtue of a comparison between the amino-acid sequences of related polypeptide chains found in contemporary organisms. A tentative partial structure is proposed for two ancestral hemoglobin polypeptide chains. Some perspectives of paleobiochemistry are outlined.

Pauling & Zuckerkandl, *Acta Chemica Scandinavica*, 1963

What is Ancestral Sequence Reconstruction (ASR)?

ASR is a computational method that reconstructs ancestral sequences based on the molecular data of present-day (extant) organisms using a model of molecular evolution

Alignment of extant molecular sequences

A model with rates of different substitutions

Phylogenetic tree

What ASR is **NOT**:

✖ A time machine

- Represent hypotheses
- Uncertainty coming from statistical inferences
- May not be the “true” ancestor

✖ A molecular clock

- ASR does not estimate divergence time
- But related tools can be used

✖ A machine learning

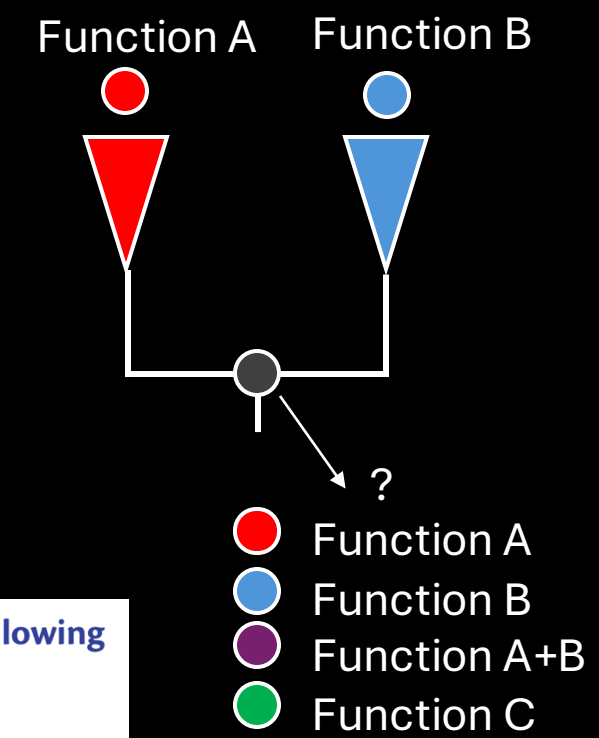
- ASR uses classical statistical models
- No training data, no pattern learning
- ML = maximum likelihood (not machine learning)

Why do we reconstruct ancestral sequences?

- What were the functions of ancestral proteins?
- Which property/function evolved first?
- Was the ancestral protein a specialist/generalist?
- Which sequence and structural changes gave rise to a new function?
- Can activities of ancestral proteins be aligned with environmental conditions?

We can get some insights into:

- Mechanisms of protein evolution
- Sequence and structural foundations of functional adaptation
- Environmental conditions in the past



Reconstruction of Nitrogenase Predecessors Suggests Origin from Maturase-Like Proteins

Amanda K. Garcia [1], Bryan Kolaczkowski [2], and Betül Kaçar [1,*]

Emergence of an Orphan Nitrogenase Protein Following Atmospheric Oxygenation

Bruno Cuevas-Zuviría,[1,2,†] Amanda K. Garcia [1,†] Alex J. Rivier,[1] Holly R. Rucker,[1] Brooke M. Carruthers,[1] and Betül Kaçar [1,*]

Kinetic Analysis Suggests Evolution of Ribosome Specificity in Modern Elongation Factor-Tus from "Generalist" Ancestors

Arindam De Tarafder [1] Narayan Prasad Parajuli [1] Soneya Majumdar [1] Betül Kaçar [2,3] and Suparna Sanyal [*,1]

Ancient nitrogenases are ATP dependent

Derek F. Harris [1], Holly R. Rucker [2], Amanda K. Garcia [2], Zhi-Yong Yang [1], Scott D. Chang [2], Hannah Feinsilber [1], Betül Kaçar [2], Lance C. Seefeldt [1]

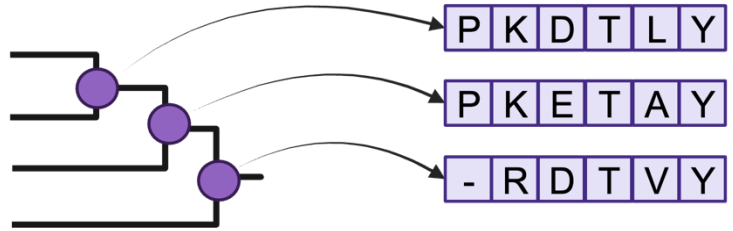
Earliest Photic Zone Niches Probed by Ancestral Microbial Rhodopsins

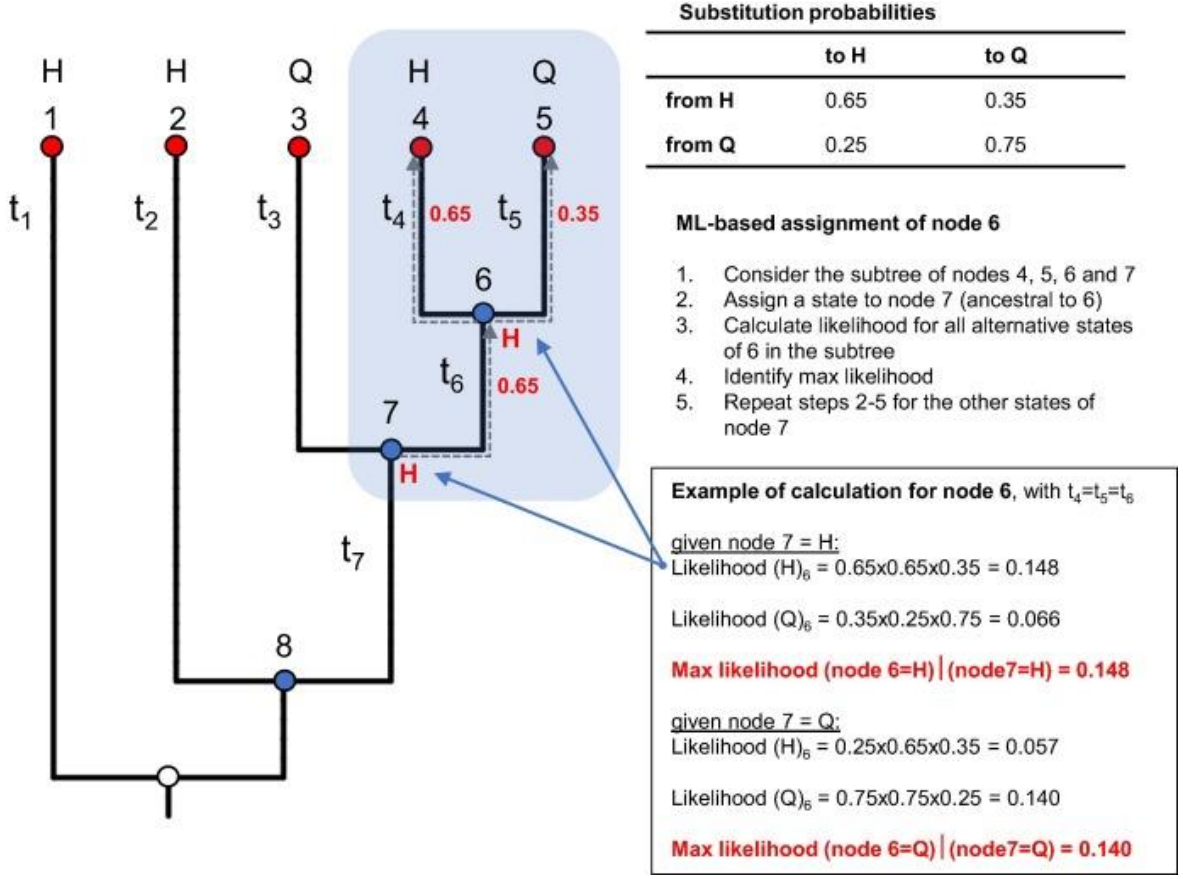
Cathryn D. Sephus,[‡1] Evrim Fer,[1,2,3] Amanda K. Garcia [1,2] Zachary R. Adam,[4,5] Edward W. Schwieterman [5,6] and Betul Kacar [*,1,2]

Evolutionary Dynamics of RuBisCO: Emergence of the Small Subunit and its Impact Through Time

Kaustubh Amritkar, Bruno Cuevas-Zuviría, Betül Kaçar ✉ Author Notes

4.Ancestral Sequence Inference





Substitution probabilities

	to H	to Q
from H	0.65	0.35
from Q	0.25	0.75

ML-based assignment of node 6

1. Consider the subtree of nodes 4, 5, 6 and 7
2. Assign a state to node 7 (ancestral to 6)
3. Calculate likelihood for all alternative states of 6 in the subtree
4. Identify max likelihood
5. Repeat steps 2-5 for the other states of node 7

Example of calculation for node 6, with $t_4=t_5=t_6$

given node 7 = H:
Likelihood $(H)_6$ = 0.65x0.65x0.35 = 0.148

Likelihood $(Q)_6$ = 0.35x0.25x0.75 = 0.066

Max likelihood (node 6=H) | (node7=H) = 0.148

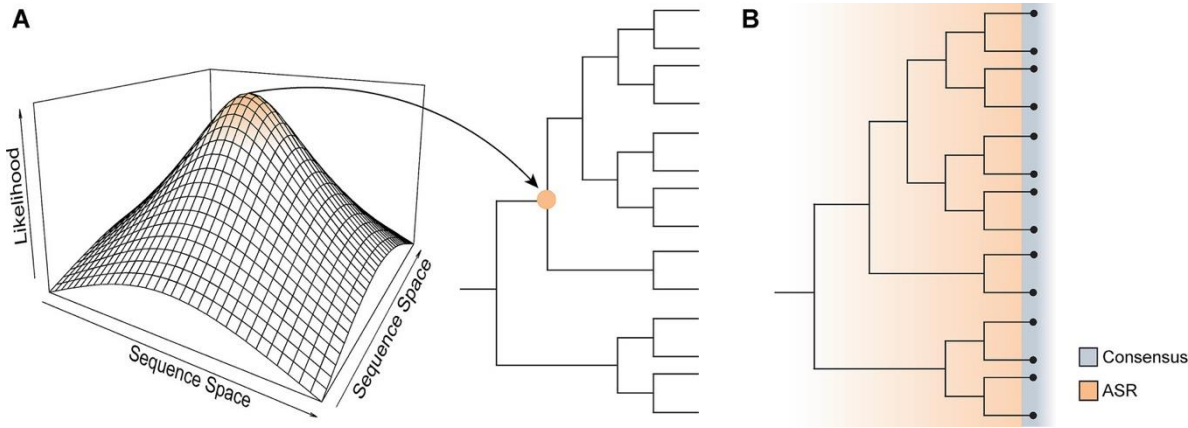
given node 7 = Q:
Likelihood $(H)_6$ = 0.25x0.65x0.35 = 0.057

Likelihood $(Q)_6$ = 0.75x0.75x0.25 = 0.140

Max likelihood (node 6=Q) | (node7=Q) = 0.140

Scossa & Fernie, *Comp. & Struct. Biotech. Journal,* 2021

ASR is different than the consensus sequence!



Garcia et al., *Methods in Mol. Bio.*, 2022

(b) Example ancestral posterior probability matrix

Node	Site	-	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Anc5	1	$p_{1,1}$	$p_{1,2}$	$p_{1,3}$	$p_{1,4}$	$p_{1,5}$	$p_{1,6}$	$p_{1,7}$	$p_{1,8}$	$p_{1,9}$	$p_{1,10}$	.....										$p_{1,21}$
⋮	⋮	⋮																				
Anc5	X	$p_{X,1}$	$p_{X,2}$	$p_{X,3}$	$p_{X,4}$	$p_{X,5}$	$p_{X,6}$	$p_{X,7}$	$p_{X,8}$	$p_{X,9}$	$p_{X,10}$	.....										$p_{X,21}$

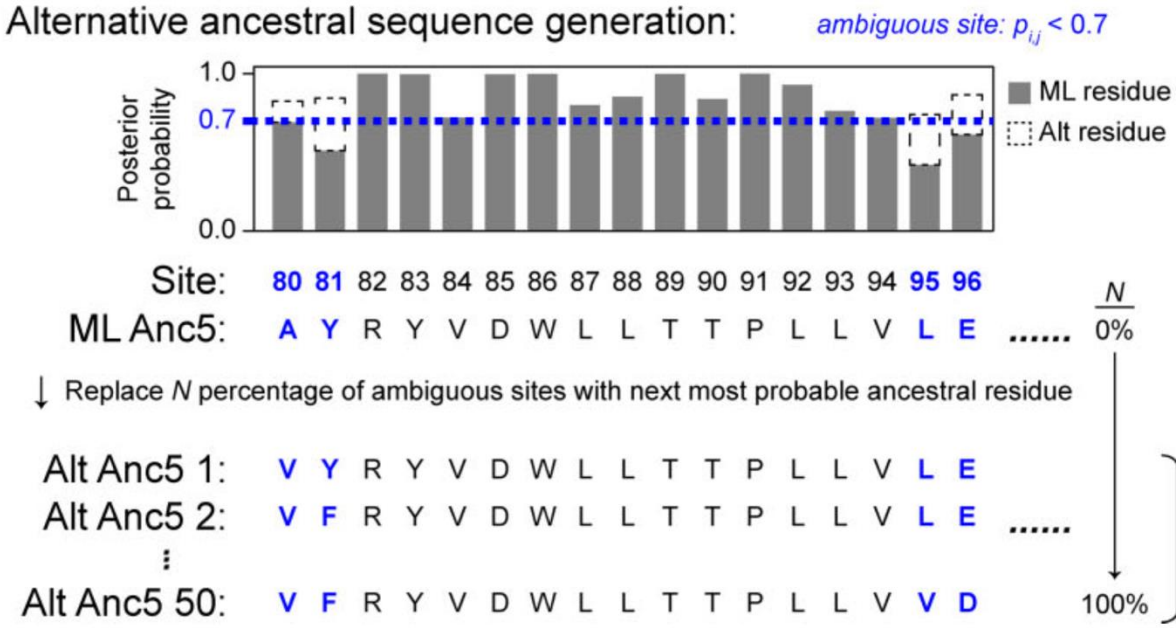
$p_{i,j}$ = posterior probability of ancestral residue, j, at site, i

Robustness of Reconstructed Ancestral Protein Functions to Statistical Uncertainty

Geeta N. Eick,[1,2] Jamie T. Bridgham,[1] Douglas P. Anderson,[1,3] Michael J. Harms,[1,3] and Joseph W. Thornton*[,4]

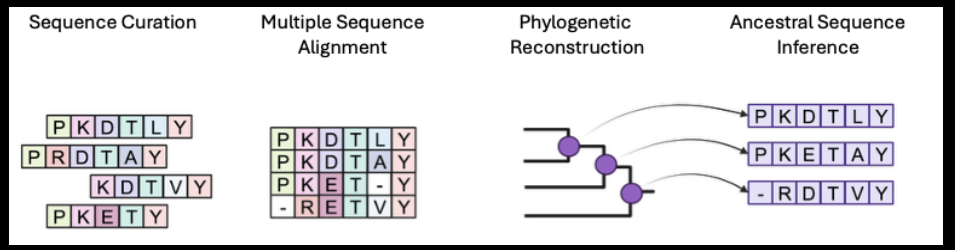
certainty about their primary sequence. Most studies to date that have addressed this question have done so by generating variants of the ML ancestral sequence, each of which contains a plausible alternate amino acid at one of the ambiguously reconstructed sites (typically defined as amino acids with a posterior probability above some arbitrary but reasonable cutoff, such as 0.2). The experimental characterization is

Eick et al., *MBE*, 2016



Sephus et al., *MBE*, 2022

Dealing with uncertainty 🧐



- Reminder: ASR is not a time machine! Most likely ancestral sequences may not be true ancestral sequence
- Uncertainties come from each step of ASR workflow: sequence dataset, alignment, evolutionary model, tree, etc…
- Carefully articulate scientific questions, conclusions given the uncertainty in ASR

Are your conclusions sensitive to ASR parameters or robust to the uncertainties?

The robustness can be tested at each step by using different methods and compare the results.

E.g., Maybe alignment generates significant uncertainty:
But we can test different alignment methods and compare the ancestral sequences resulted from each
If they all yield proteins with similar experimental properties, we can be confident in our conclusions

E.g., Maybe evolutionary model generates uncertainty:
We can test second best evolutionary model and, again, compare the ancestral sequences resulted from each
If both result ancestral proteins with properties, we can be confident in our conclusions

We have the ancestral sequences… What is next?

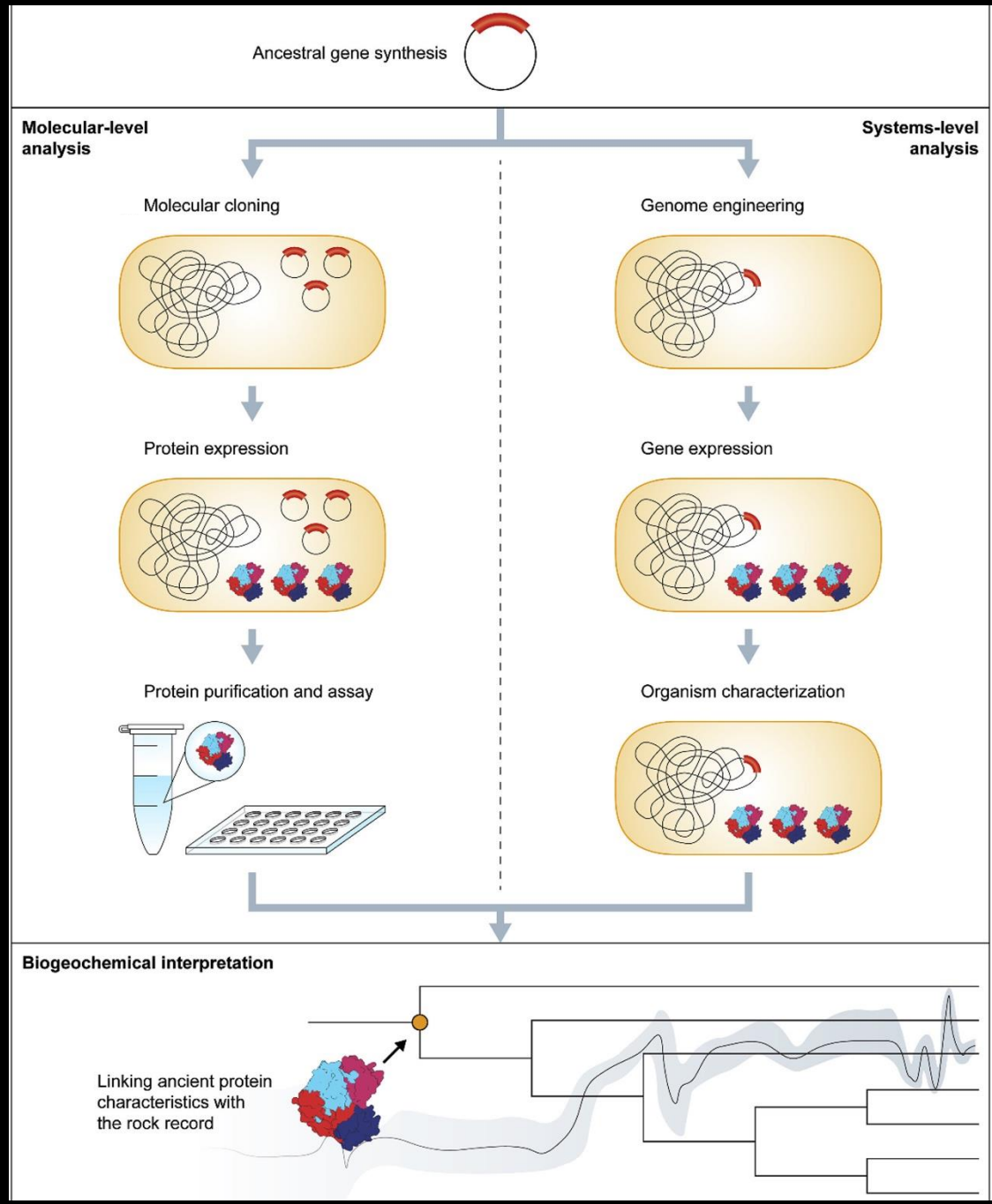
Important considerations moving toward next steps:

- Which model organism to use to capture property of interest?
- Codon optimization to improve translation efficiency in expression system
- Which genes to replace? Location of genetic insertion?
- *In vivo* or *in vitro* characterization?

Impact of cellular interactions and environments…

Ancient proteins presumably adapted toward environments, interactions partners, and substrates.

How to account for mismatches between ancient proteins and modern environments?



Garcia & Kaçar, *Free Radical Bio. & Med.*, 2019

Any question?

Acknowledgement:

- Some slides are modified from Dr. Amanda Garcia’s ASR Workshop presentation

Tutorial time

- ASR