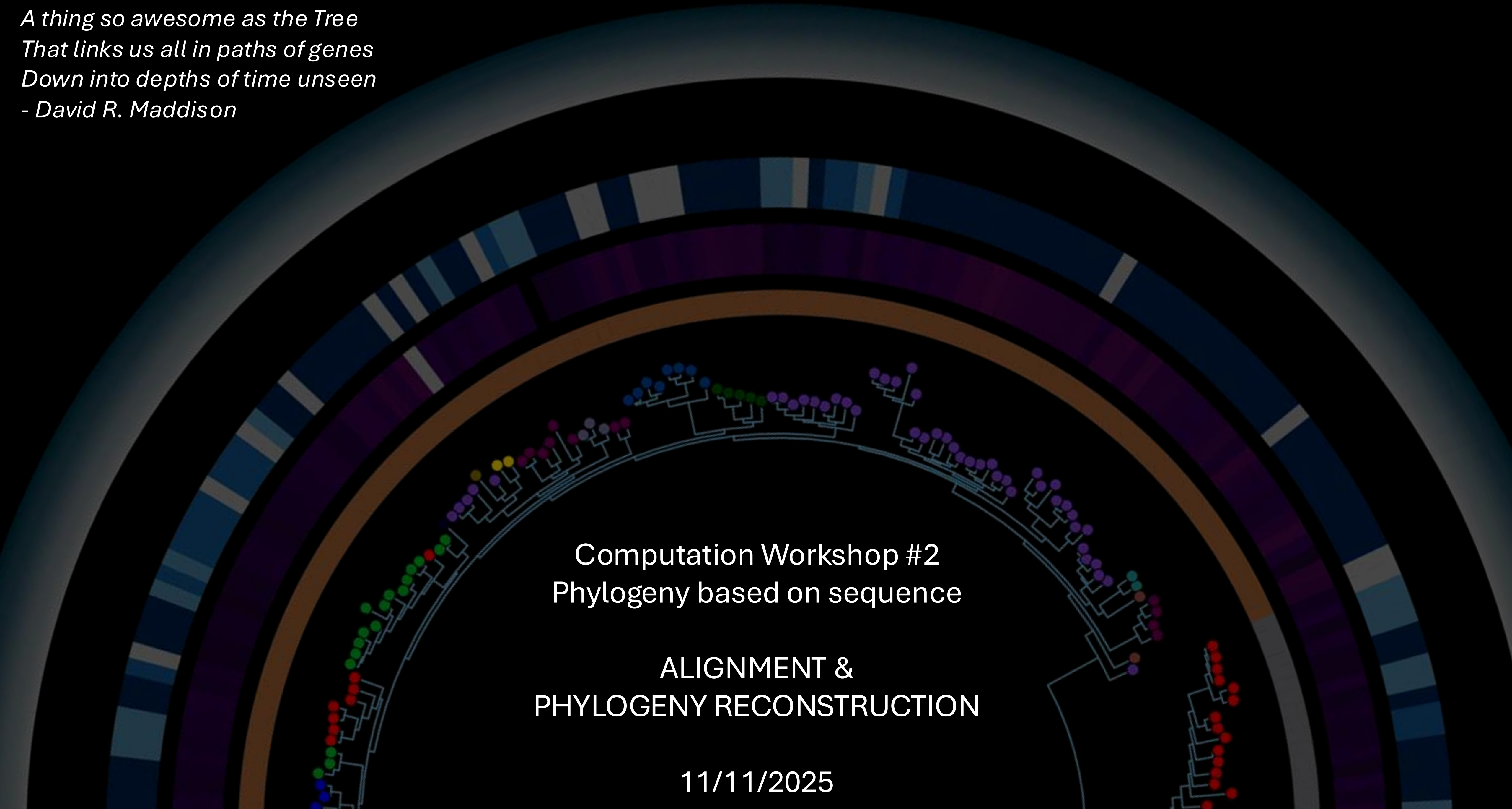
I think that I shall never see
A thing so awesome as the Tree
That links us all in paths of genes
Down into depths of time unseen
- David R. Maddison

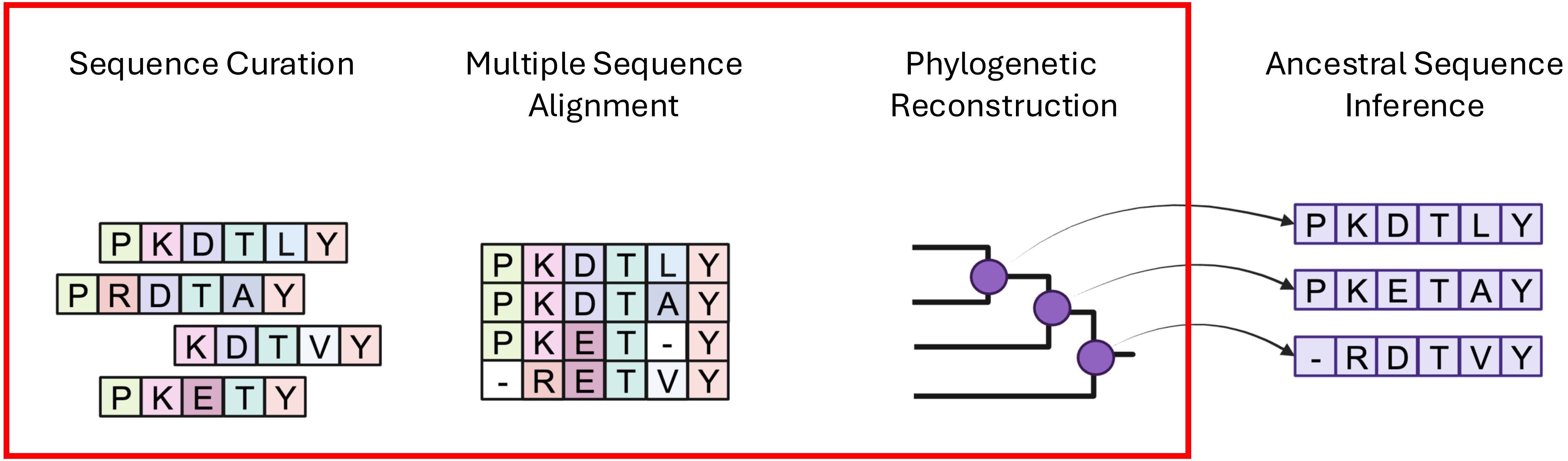


Computation Workshop #2
Phylogeny based on sequence

ALIGNMENT &
PHYLOGENY RECONSTRUCTION

11/11/2025

ASR Workflow



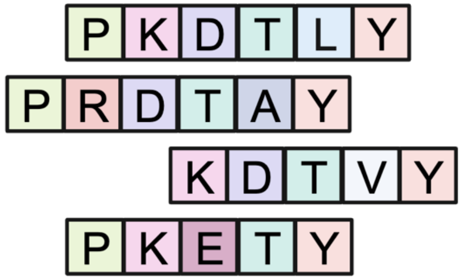
Sequence Curation

Multiple Sequence Alignment

Phylogenetic Reconstruction

Ancestral Sequence Inference

1. Sequence Curation



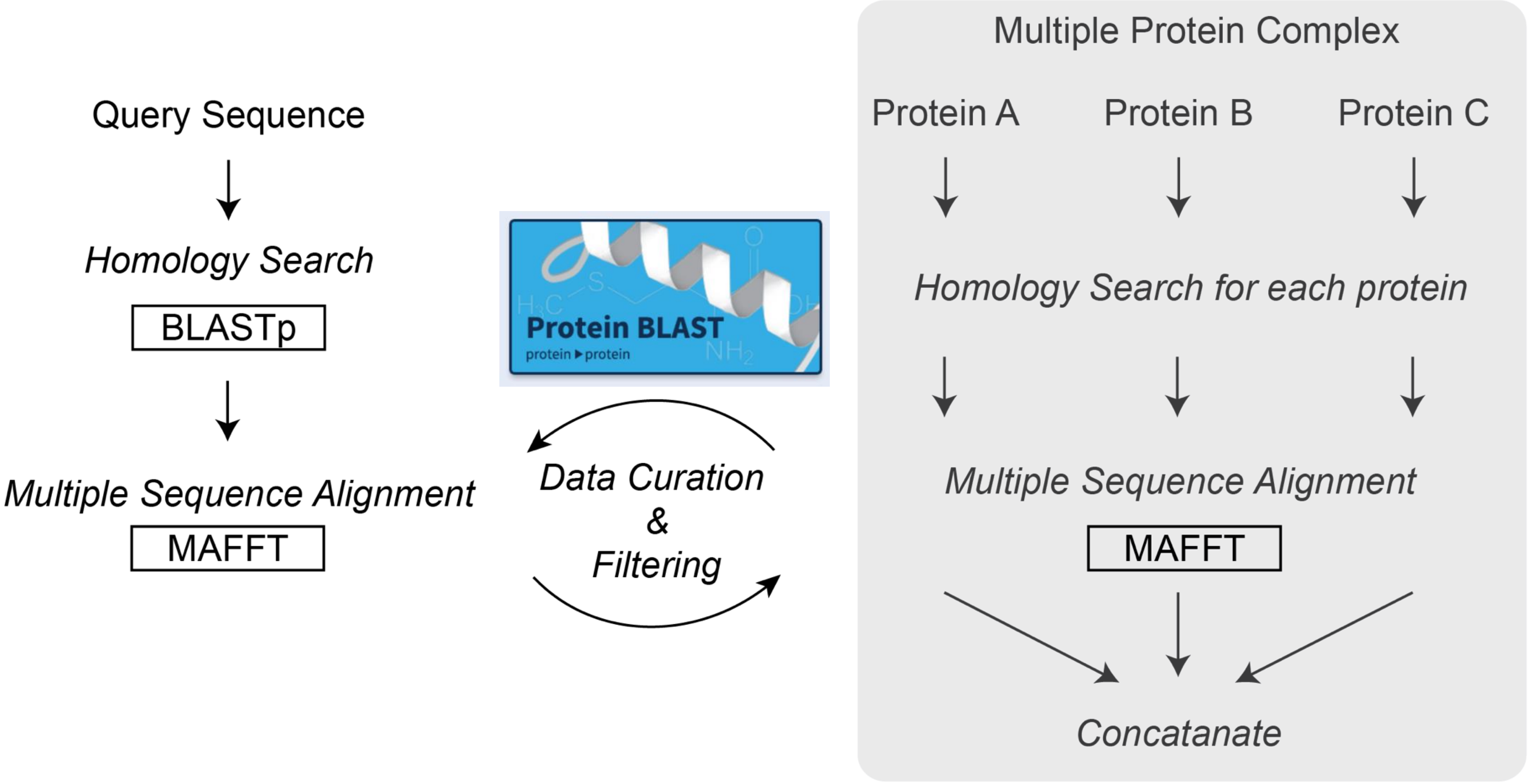
```
 PKDTLY
PRDTAY
   KDTVY
 PKETY
```

- Collection of homolog protein sequences from different organisms
- The most important step!

- Need to consider:
  - Which ancestor I am interested in?
  - Which group of modern organisms should I collect?
  - What can I use as outgroup?
  - What should be the size of the final dataset?
  - Single gene or multiple genes?

1. Sequence Curation

```
 P K D T L Y
P R D T A Y
   K D T V Y
 P K E T Y
```



Query Sequence

Homology Search

BLASTp

Multiple Sequence Alignment

MAFFT

Data Curation & Filtering

Multiple Protein Complex

Protein A

Protein B

Protein C

Homology Search for each protein

Multiple Sequence Alignment

MAFFT

Concatanate

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

2. Multiple Sequence Alignment (MSA)

P	K	D	T	L	Y
P	K	D	T	A	Y
P	K	E	T	-	Y
-	R	E	T	V	Y

Goal: introduce gaps into sequences so that columns of alignment contain character states that are homologous.

MAFFT

PRANK

MUSCLE

CLUSTAL

2. Multiple Sequence Alignment (MSA)

P	K	D	T	L	Y
P	K	D	T	A	Y
P	K	E	T	-	Y
-	R	E	T	V	Y

- MSA algorithms aim to evaluate multiple possible alignments and select the one with the <u>highest score</u>.
- Aligning identical characters increases the score, while aligning different characters reduces the score.
- The extent of the score parameters depends on the substitution matrix involved.

Alignment 1 ✓

M	L	T	T	T	C	
M	L	A	-	-	C	
+5	+5	-3	-4	-1	+5	= 7

Alignment 2

M	L	T	T	T	C	
M	L	-	A	-	C	
+5	+5	-4	-3	-4	+5	= 4

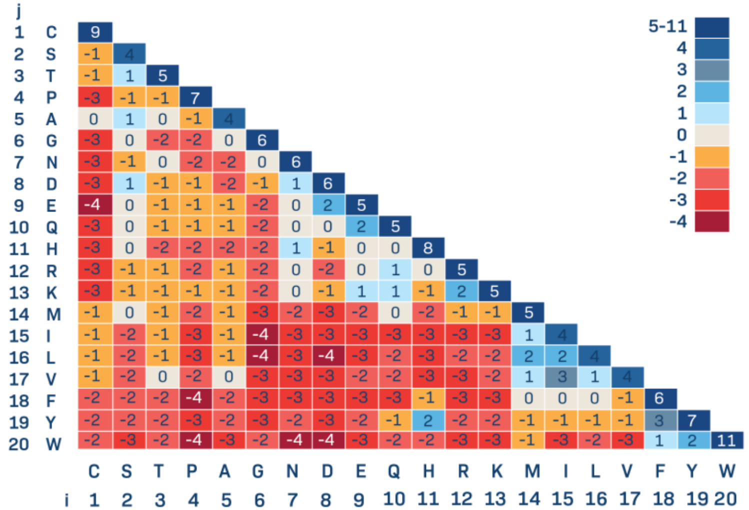
Scoring Parameters

Match= +5

Mismatch= -3

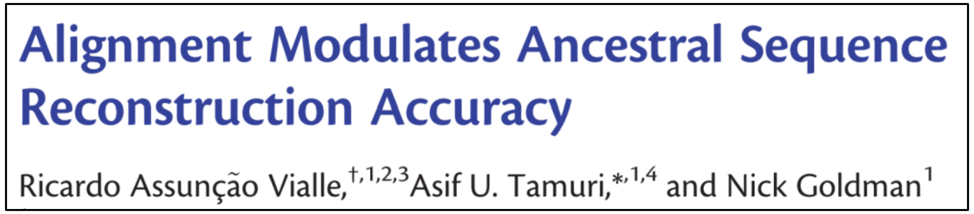
Gap open= -4

Gap extension= -1



j																					
1	C	9																			
2	S	-1	4																		
3	T	-1	1	5																	
4	P	-3	-1	-1	7																
5	A	0	1	0	-1	4															
6	G	-3	0	-2	-2	0	6														
7	N	-3	-1	0	-2	-2	0	6													
8	D	-3	1	-1	-1	-2	-1	1	6												
9	E	-4	0	-1	-1	-1	-2	0	2	5											
10	Q	-3	0	-1	-1	-1	-2	0	0	2	5										
11	H	-3	0	-2	-2	-2	-2	1	-1	0	0	8									
12	R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
13	K	-3	-1	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
14	M	-1	0	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
15	I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
16	L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
17	V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
18	F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
19	Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
20	W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
	i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

BLOSUM Substitution Matrix



Alignment Modulates Ancestral Sequence Reconstruction Accuracy

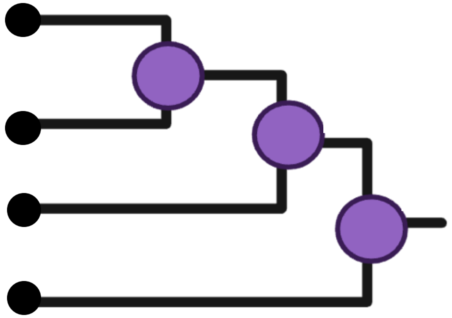
Ricardo Assunção Vialle,[†,1,2,3] Asif U. Tamuri,[*,1,4] and Nick Goldman[1]

A source of ASR inaccuracy!

Tutorial time

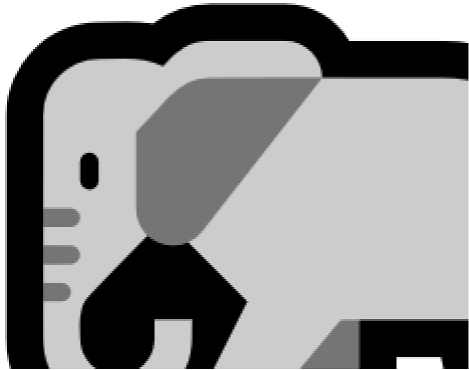
- Running MAFFT

3.Phylogeny Reconstruction

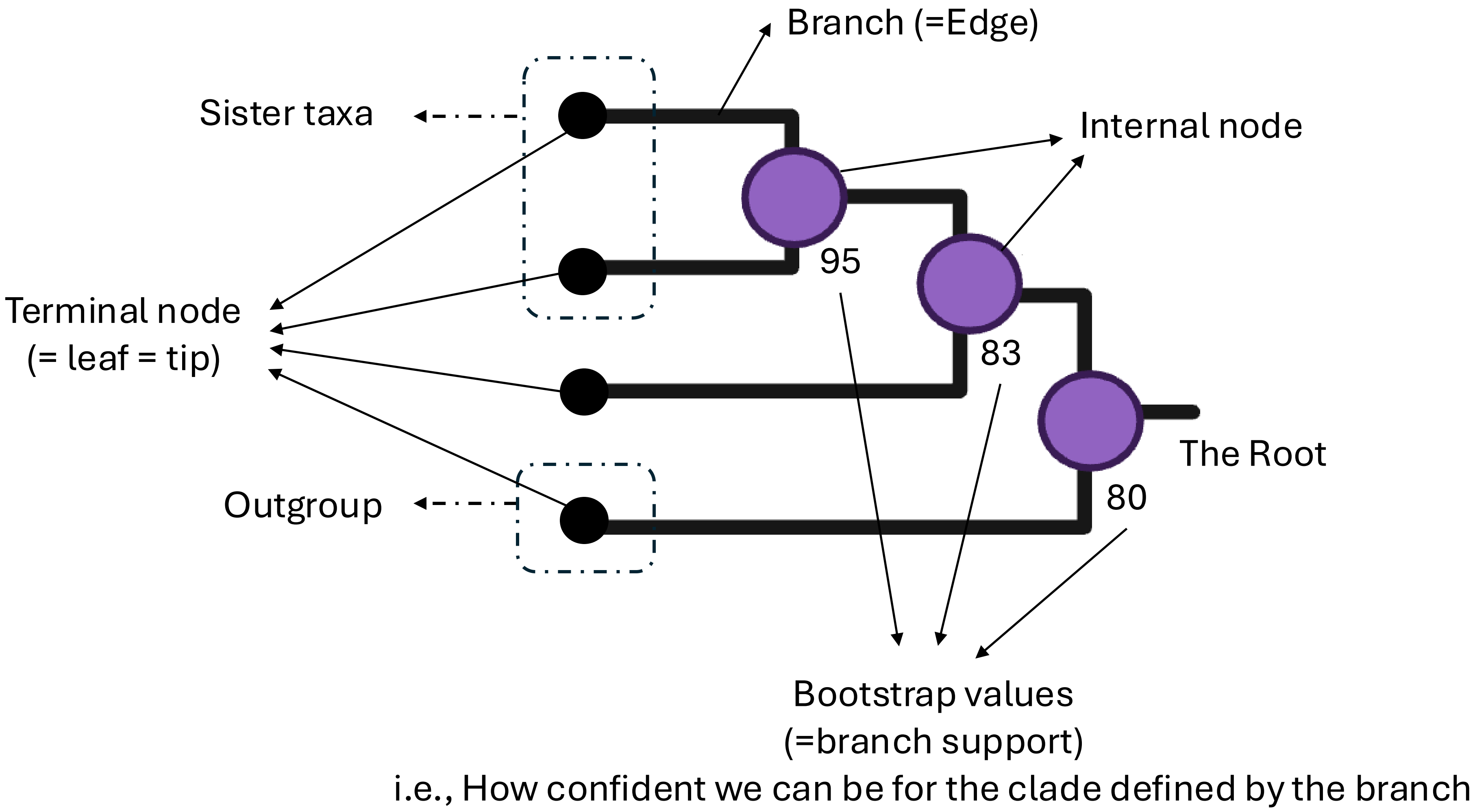


A phylogenetic tree:

- Describes evolutionary relationships among a group of organisms based on their molecular sequences.
- Represents a model of evolutionary history depicted by ancestor-descendant relationships between organisms at different level of relatedness.



Topology of A Phylogenetic Tree

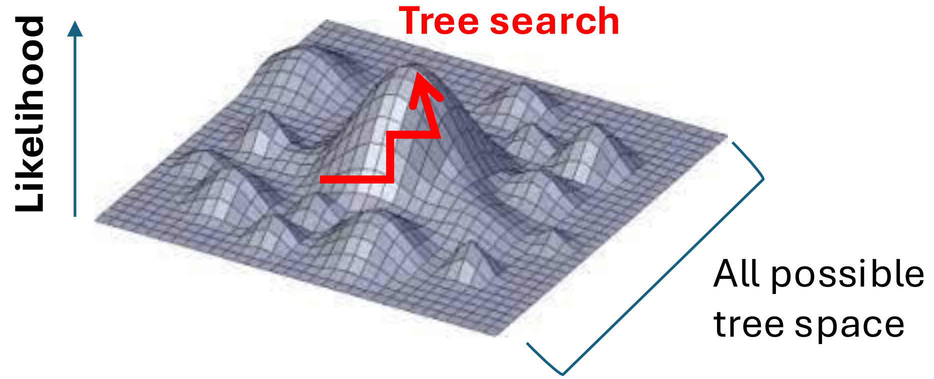


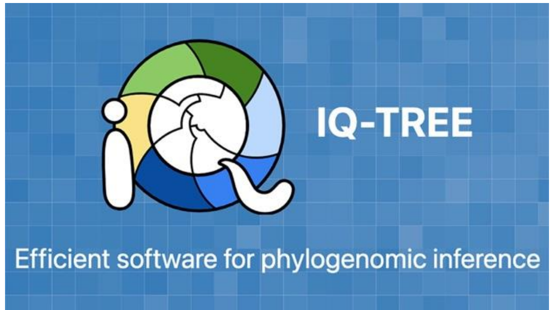
Construction of Phylogenetic Trees

Alignment of extant molecular sequences

A model with rates of different substitutions

- Distance based = Neighbor-joining
- Character based = Maximum parsimony
- Statistical = Maximum Likelihood, Bayesian

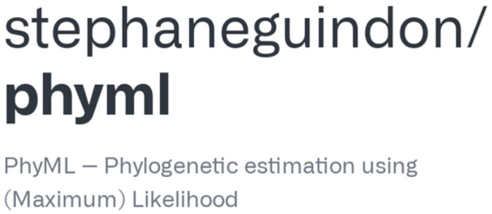



IQ-TREE

Efficient software for phylogenomic inference

amkozlov/**raxml-ng**

RAxML Next Generation: faster, easier-to-use and more flexible


stephaneguindon/
phyml

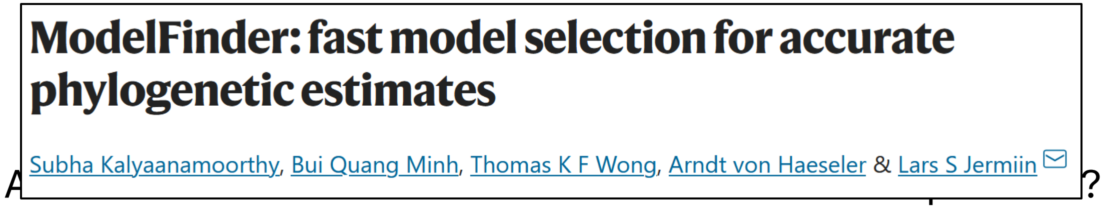
PhyML – Phylogenetic estimation using (Maximum) Likelihood

PAML:
Phylogenetic Analysis
by Maximum Likelihood

Tutorial time

- Running IQTREE

Why do we need an evolutionary model?

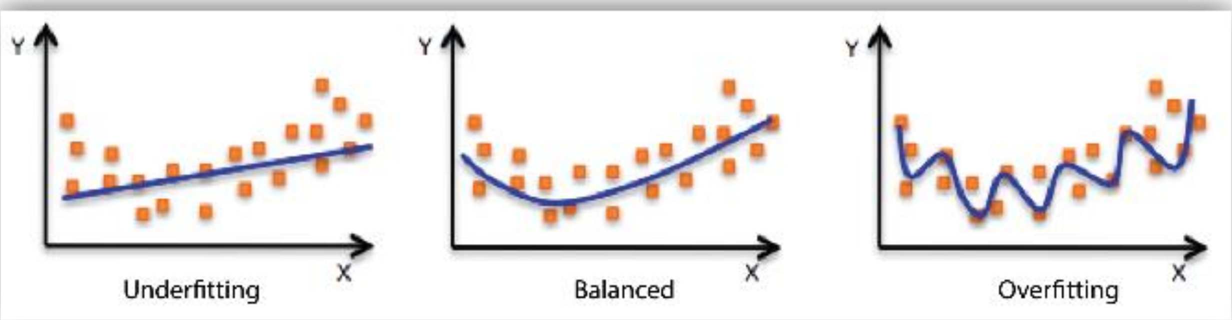


ModelFinder: fast model selection for accurate phylogenetic estimates

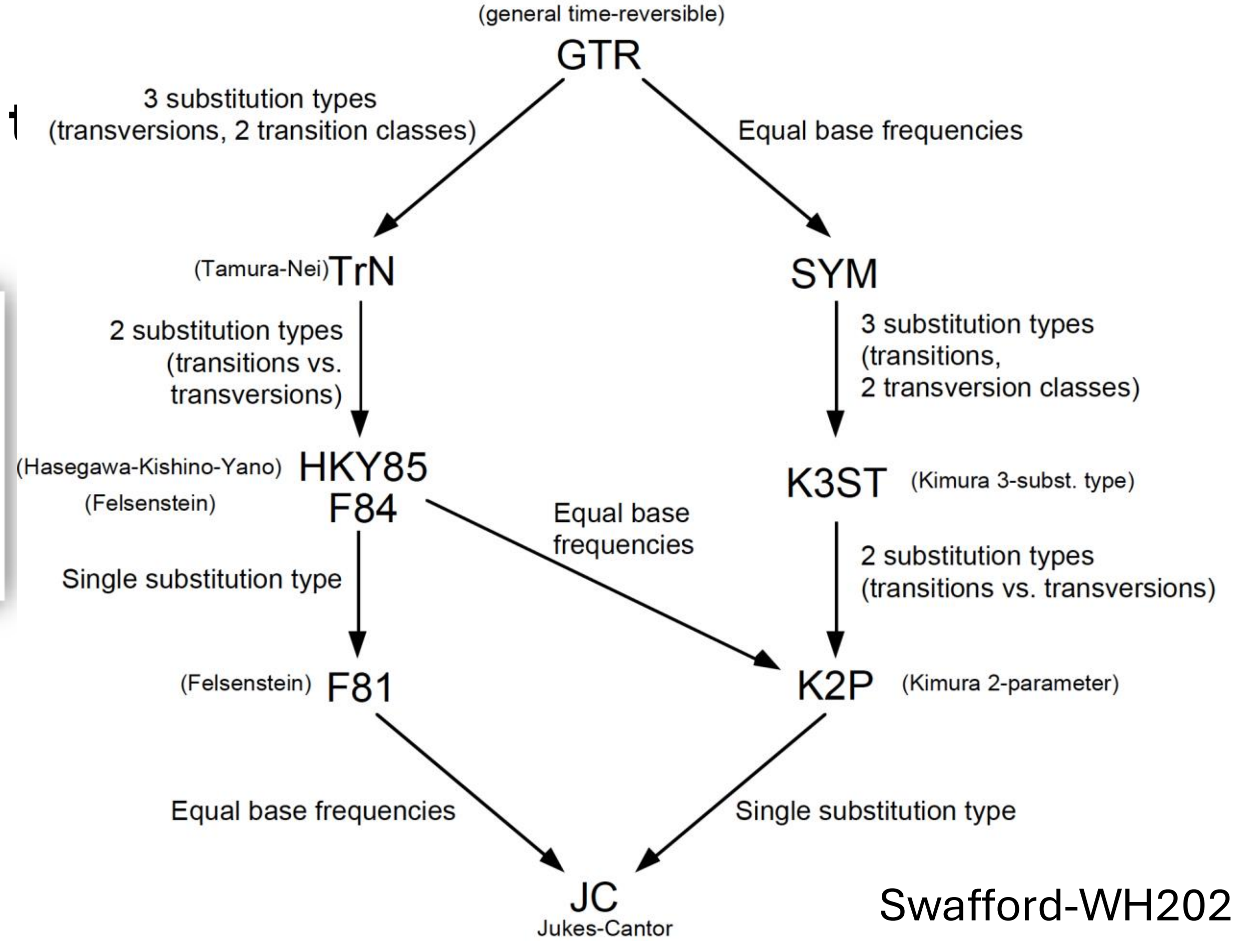
Subha Kalyaanamoorthy, Bui Quang Minh, Thomas K F Wong, Arndt von Haeseler & Lars S Jermiin

A ?

Evolutionary models account for t



- Underfitting models oversimplify molecular evolution
- Overfitting models has too much parameter, takes too much time to calculate likelihood



(general time-reversible) GTR

3 substitution types (transversions, 2 transition classes)

Equal base frequencies

(Tamura-Nei) TrN

SYM

2 substitution types (transitions vs. transversions)

3 substitution types (transitions, 2 transversion classes)

(Hasegawa-Kishino-Yano) HKY85

(Felsenstein) F84

K3ST (Kimura 3-subst. type)

Single substitution type

Equal base frequencies

2 substitution types (transitions vs. transversions)

(Felsenstein) F81

K2P (Kimura 2-parameter)

Equal base frequencies

Single substitution type

JC

Jukes-Cantor

Swafford-WH2025

Jukes & Cantor (1969)

$$Q = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix}$$

Kimura (1980)

$$Q = \begin{pmatrix} -1 & 1/(\kappa+2) & \kappa/(\kappa+2) & 1/(\kappa+2) \\ 1/(\kappa+2) & -1 & 1/(\kappa+2) & \kappa/(\kappa+2) \\ \kappa/(\kappa+2) & 1/(\kappa+2) & -1 & 1/(\kappa+2) \\ 1/(\kappa+2) & \kappa/(\kappa+2) & 1/(\kappa+2) & -1 \end{pmatrix}$$

Hasegawa, Kishino, and Yano (1985)

$$Q = \begin{pmatrix} - & \pi_{\mathrm{C}} & \kappa\pi_{\mathrm{G}} & \pi_{\mathrm{T}} \\ \pi_{\mathrm{A}} & - & \pi_{\mathrm{G}} & \kappa\pi_{\mathrm{T}} \\ \kappa\pi_{\mathrm{A}} & \pi_{\mathrm{C}} & - & \pi_{\mathrm{T}} \\ \pi_{\mathrm{A}} & \kappa\pi_{\mathrm{C}} & \pi_{\mathrm{G}} & - \end{pmatrix}\mu$$

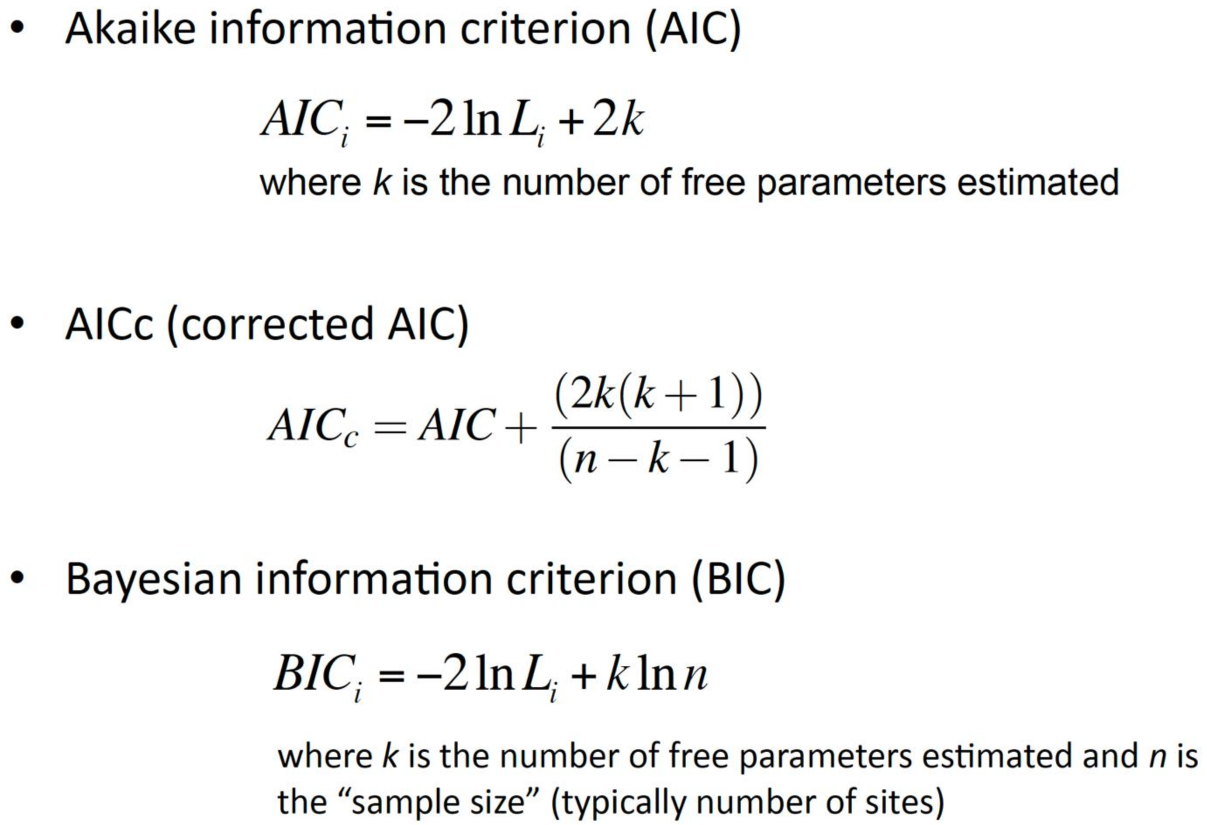
GTR (Tavare, 1986)

$$Q = \begin{pmatrix} - & r_{AC}\pi_{\mathrm{C}} & r_{AG}\pi_{\mathrm{G}} & r_{AT}\pi_{\mathrm{T}} \\ r_{AC}\pi_{\mathrm{A}} & - & r_{CG}\pi_{\mathrm{G}} & r_{CT}\pi_{\mathrm{T}} \\ r_{AG}\pi_{\mathrm{A}} & r_{CG}\pi_{\mathrm{C}} & - & \pi_{\mathrm{T}} \\ r_{AT}\pi_{\mathrm{A}} & r_{CT}\pi_{\mathrm{C}} & \pi_{\mathrm{G}} & - \end{pmatrix}\mu$$

There are two criteria for choosing the best model:

- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)

Both try to find the balance between underfitting and overfitting models



- Akaike information criterion (AIC)

$$AIC_i = -2\ln L_i + 2k$$

where *k* is the number of free parameters estimated

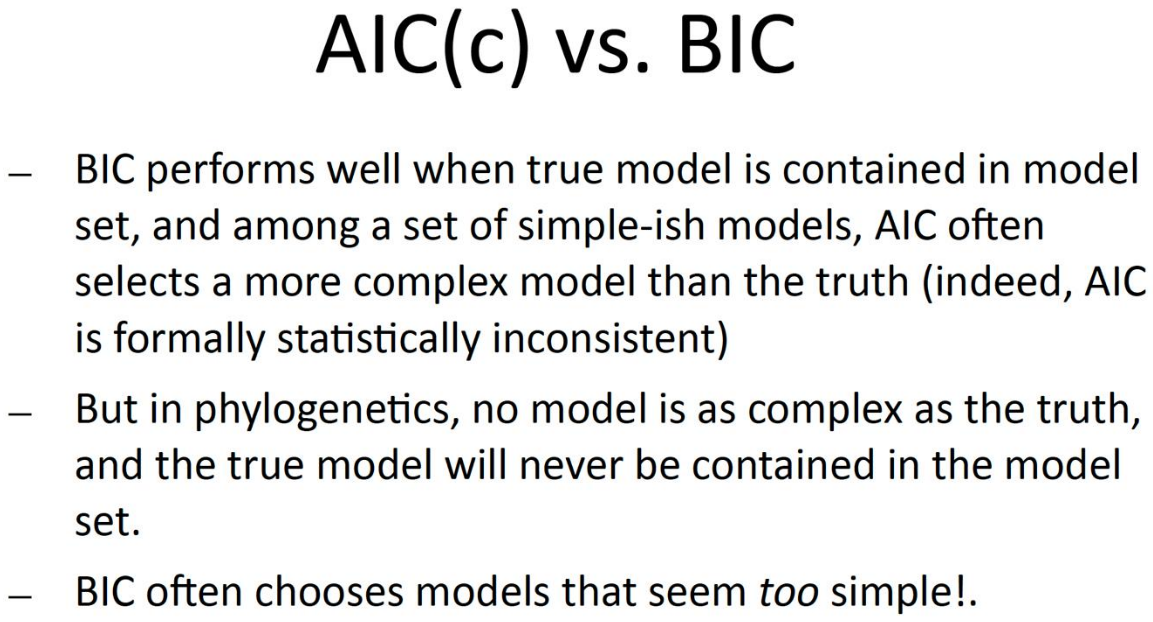
- AICc (corrected AIC)

$$AIC_c = AIC + \frac{(2k(k+1))}{(n-k-1)}$$

- Bayesian information criterion (BIC)

$$BIC_i = -2\ln L_i + k\ln n$$

where *k* is the number of free parameters estimated and *n* is the “sample size” (typically number of sites)



AIC(c) vs. BIC

– BIC performs well when true model is contained in model set, and among a set of simple-ish models, AIC often selects a more complex model than the truth (indeed, AIC is formally statistically inconsistent)

– But in phylogenetics, no model is as complex as the truth, and the true model will never be contained in the model set.

– BIC often chooses models that seem *too* simple!.

Tutorial time

- Navigating output

Any question?

Acknowledgement:

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