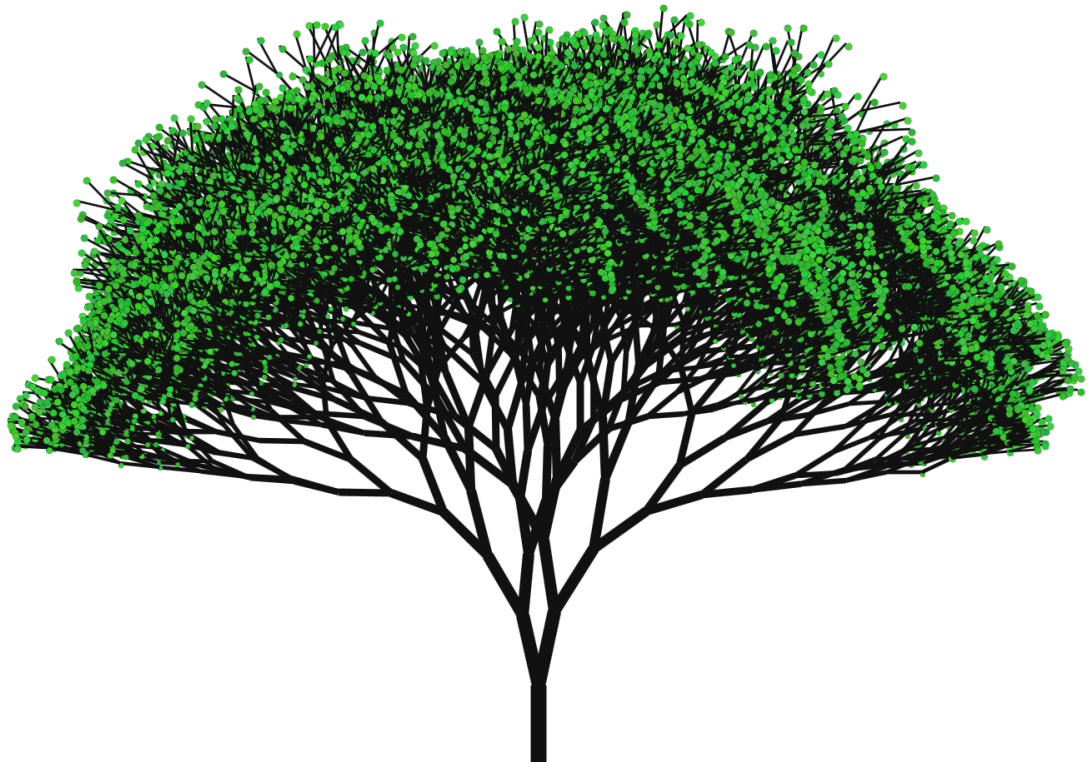


Predicting, exploring, and synthesising ancestral sequences using Graphical Representation of Ancestral Sequence Predictions (GRASP)

Gabe Foley

School of Chemistry and Molecular Biosciences

The University of Queensland

10/12/2019

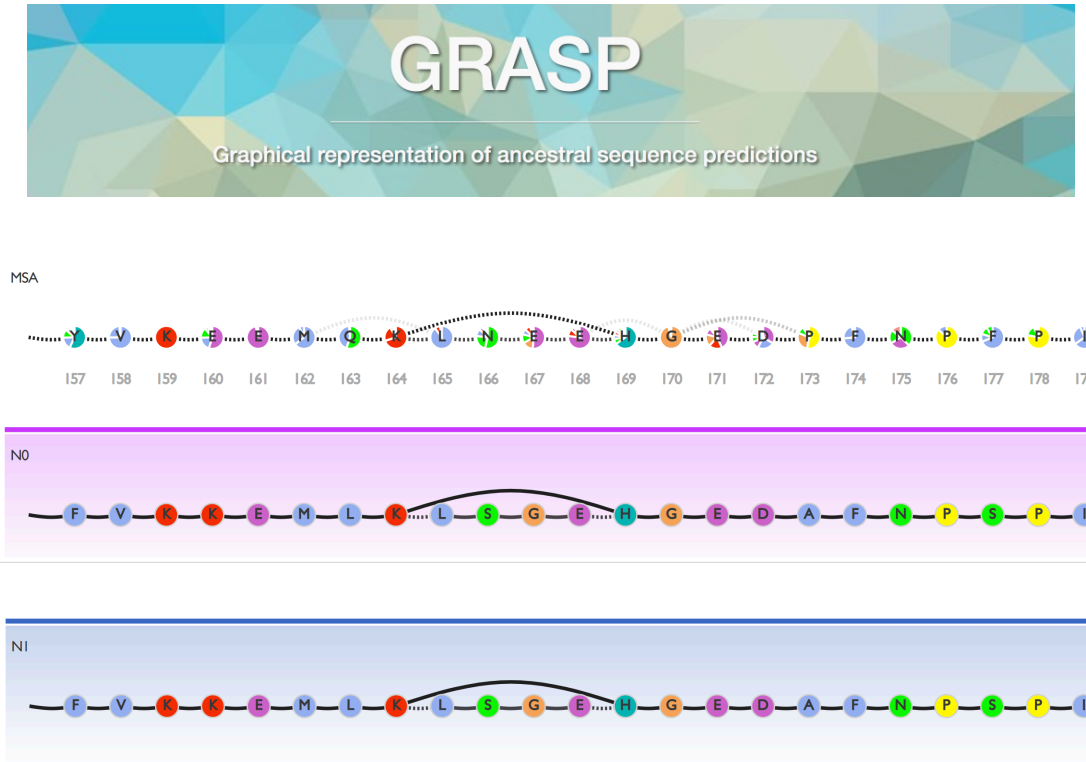
Overview

Ancestral Sequence Reconstruction (ASR)

- What is it?
- Why use it?
- Current restrictions on data set size

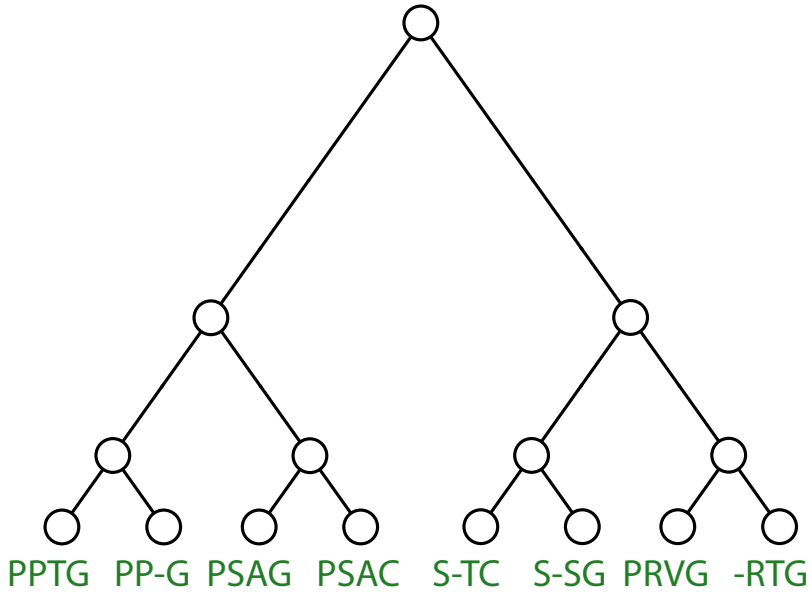
Graphical Representation of Ancestral Sequence Predictions (GRASP)

- Enables much greater data set sizes
- Valid ancestral predictions
- Allows for novel types of ancestors



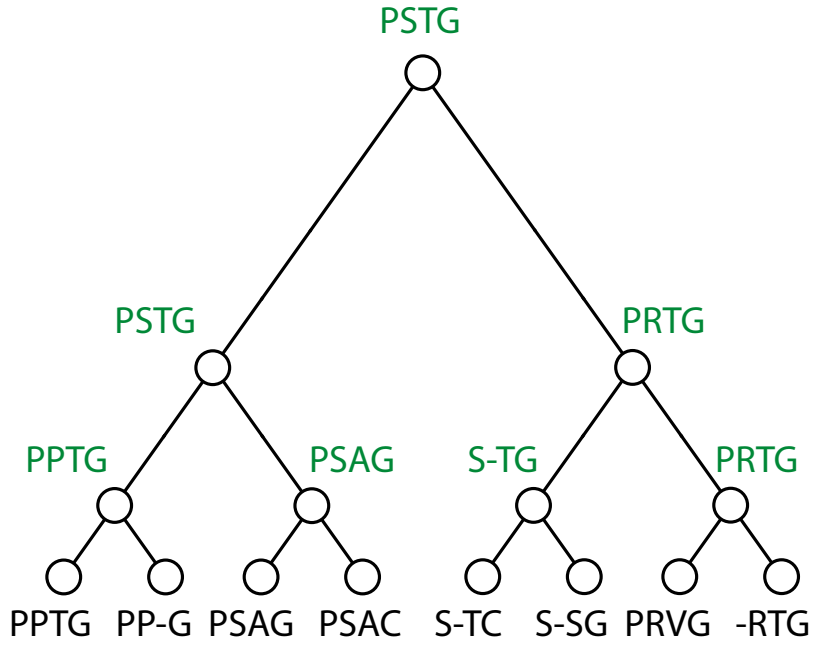
What is ancestral sequence reconstruction?

- Using the information in modern day biological sequences to infer what their ancestors looked like



What is ancestral sequence reconstruction?

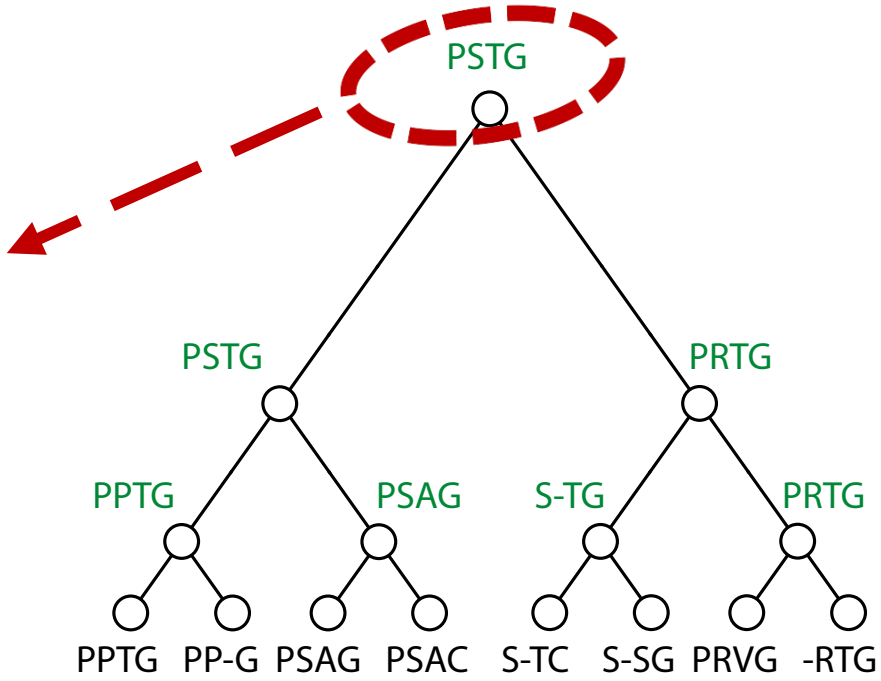
- Using the information in modern day biological sequences to infer what their ancestors looked like



What is ancestral sequence reconstruction?

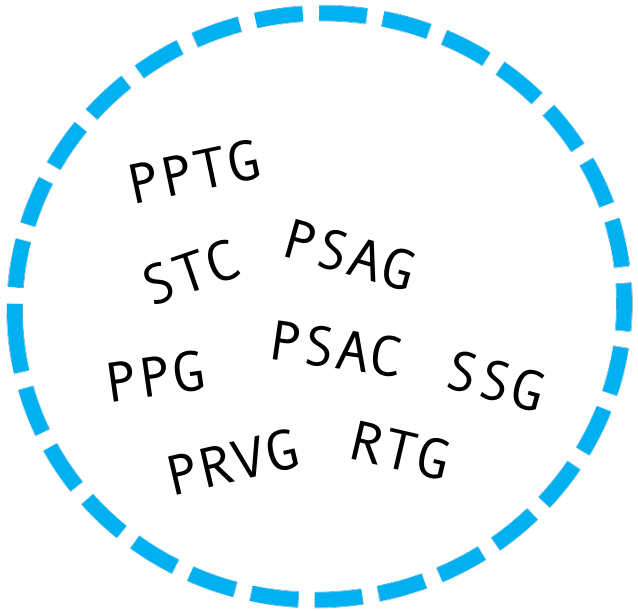
- Using the information in modern day biological sequences to infer what their ancestors looked like

- Ancestral sequences can be ‘resurrected’ – synthesised and studied alongside modern day proteins

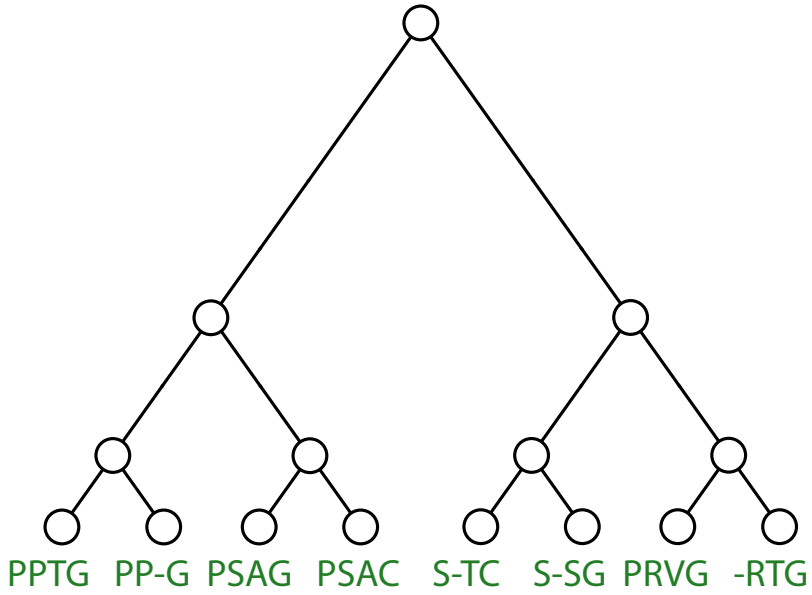


Ancestral sequence reconstruction steps

1. Collect sequences



PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

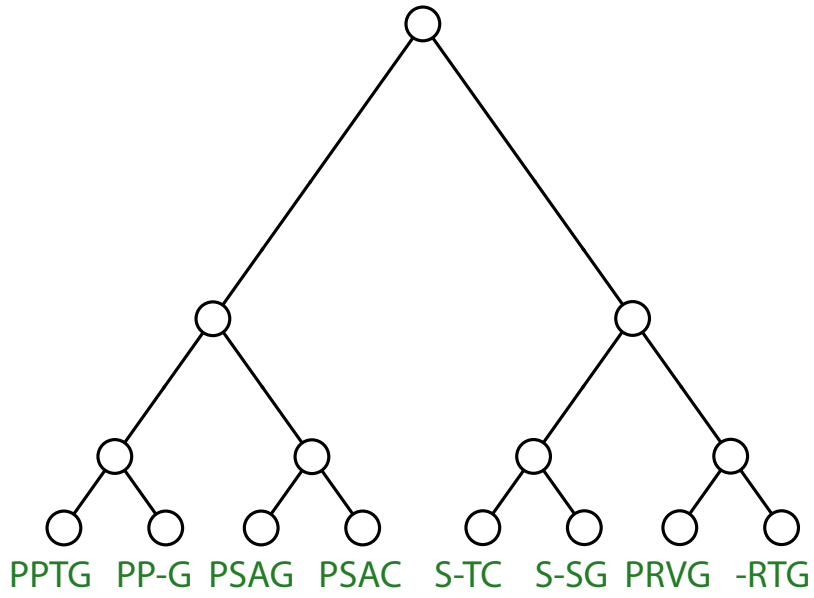


Ancestral sequence reconstruction steps

2. Align sequences

PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

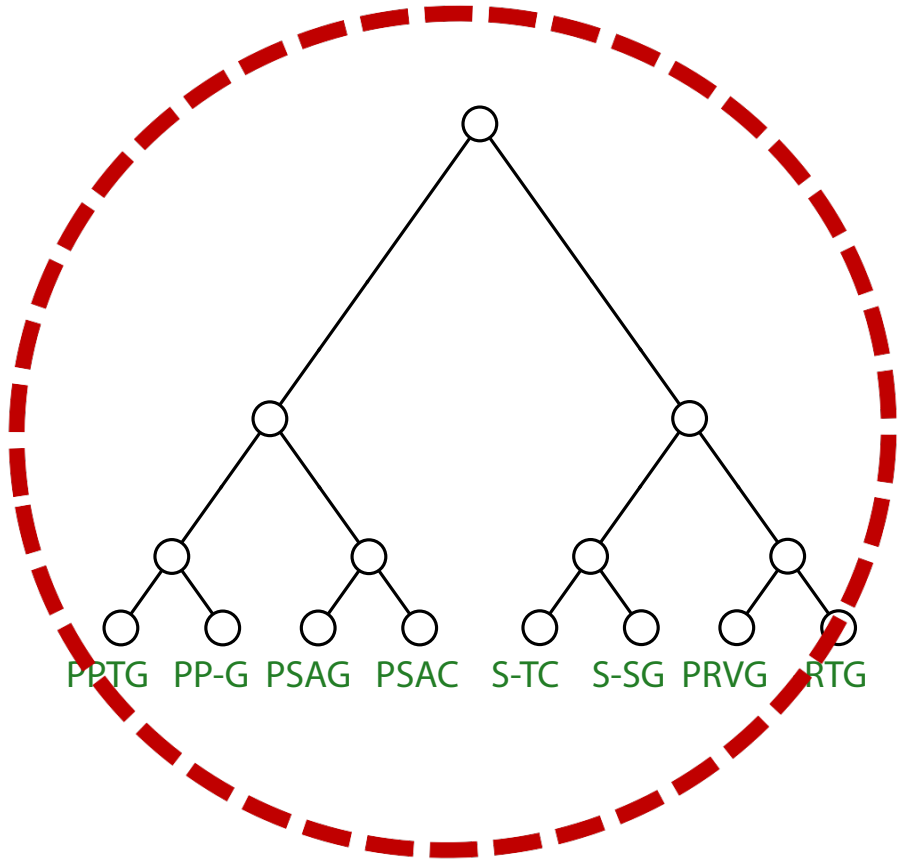


Ancestral sequence reconstruction steps

3. Infer phylogenetic tree

PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

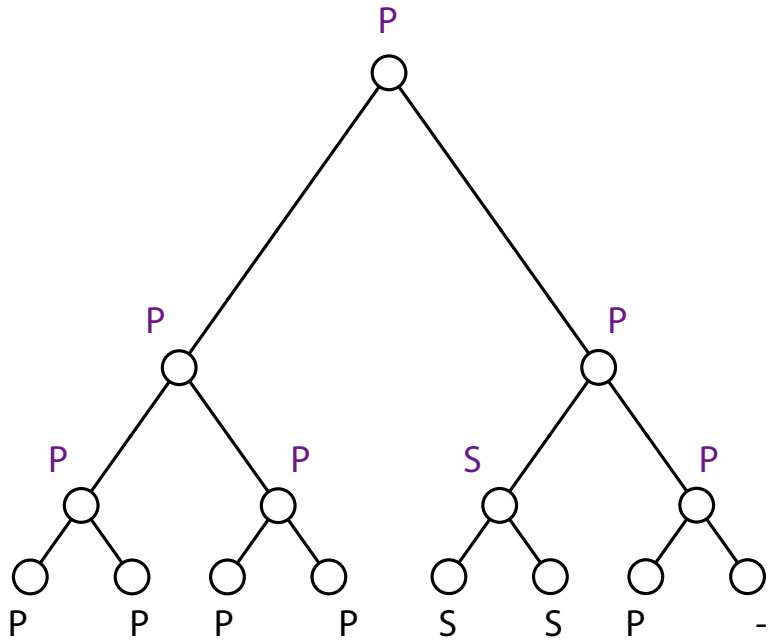


Ancestral sequence reconstruction steps

4. Infer ancestors for individual columns

PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

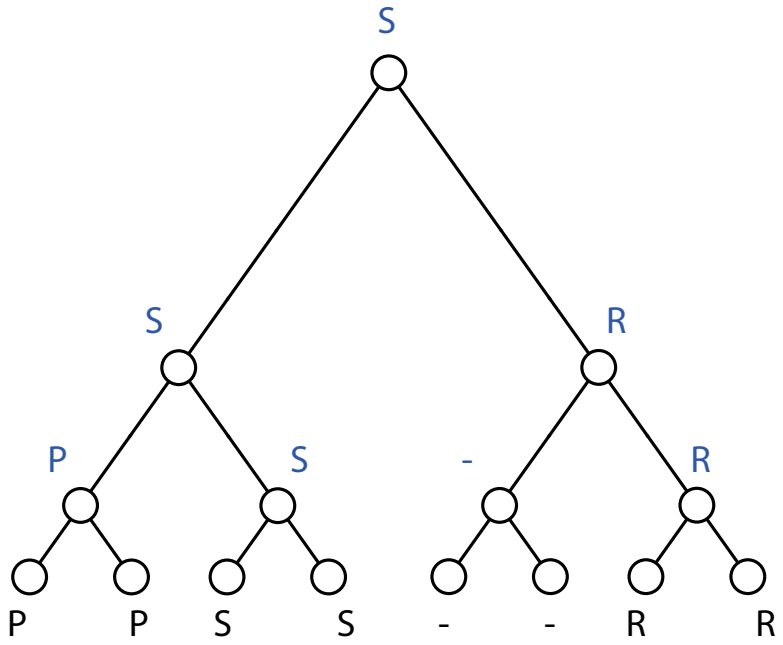


Ancestral sequence reconstruction steps

4. Infer ancestors for individual columns

PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

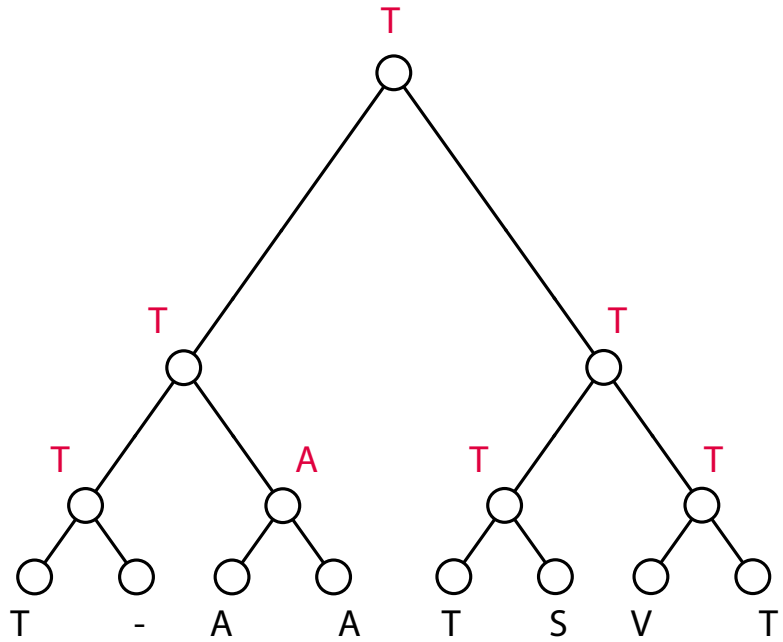


Ancestral sequence reconstruction steps

4. Infer ancestors for individual columns

PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

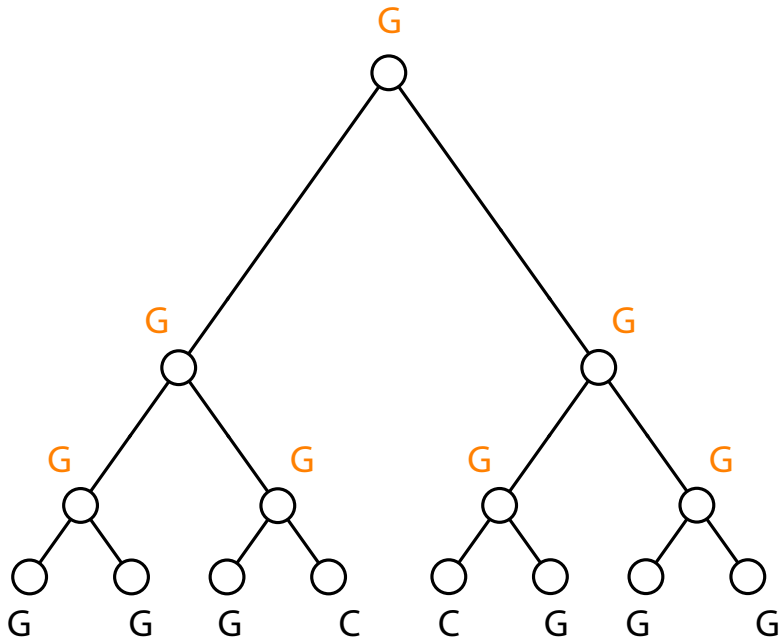


Ancestral sequence reconstruction steps

4. Infer ancestors for individual columns

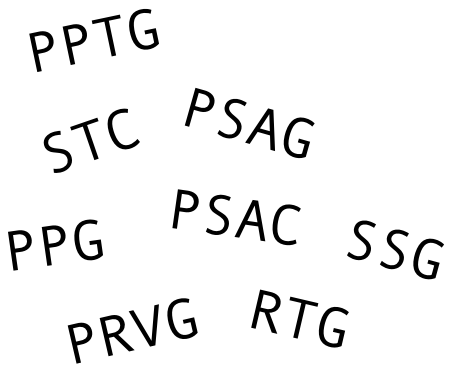
PPTG
STC PSAG
PPG PSAC SSG
PRVG RTG

PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG

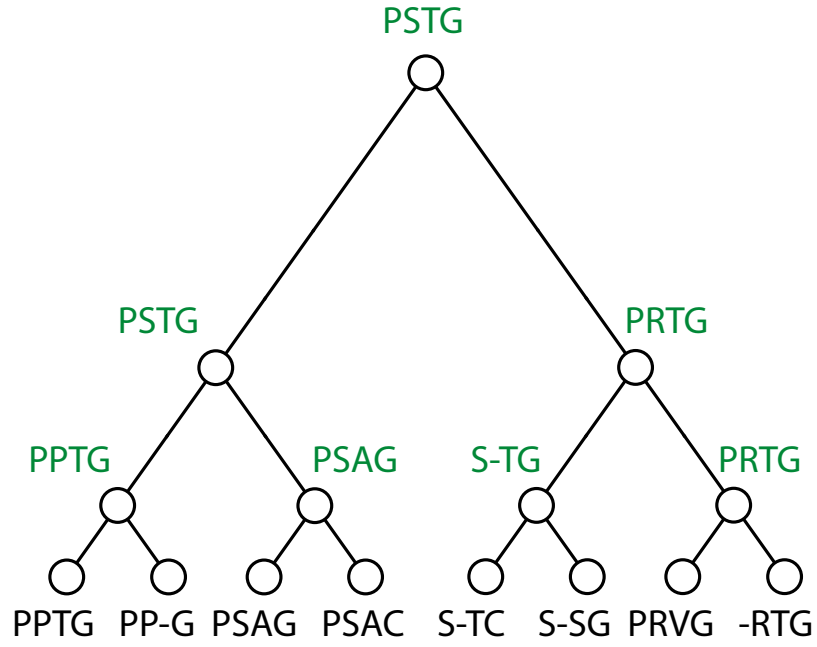


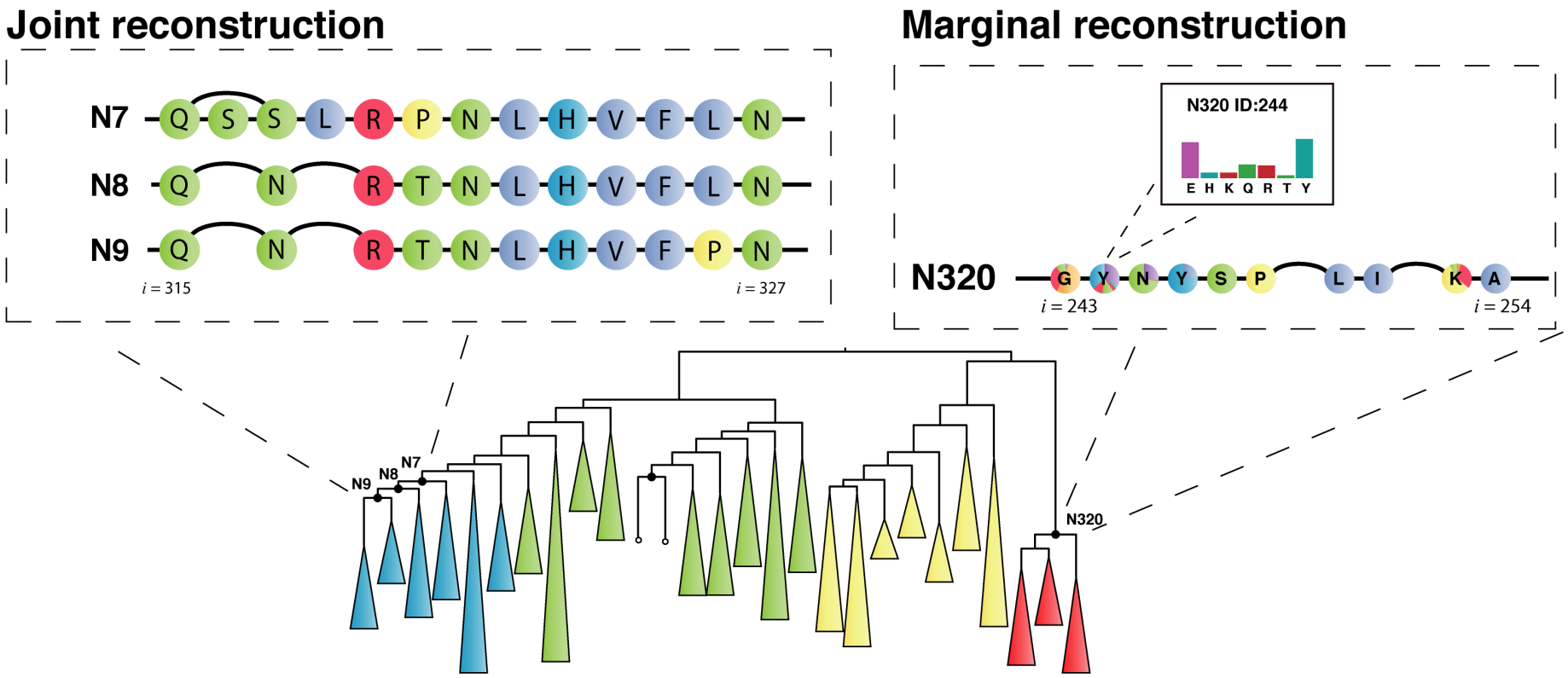
Ancestral sequence reconstruction steps

5. Concatenate predictions into a complete sequence



PPTG
PP-G
PSAG
PSAC
S-TC
S-SG
PRVG
-RTG





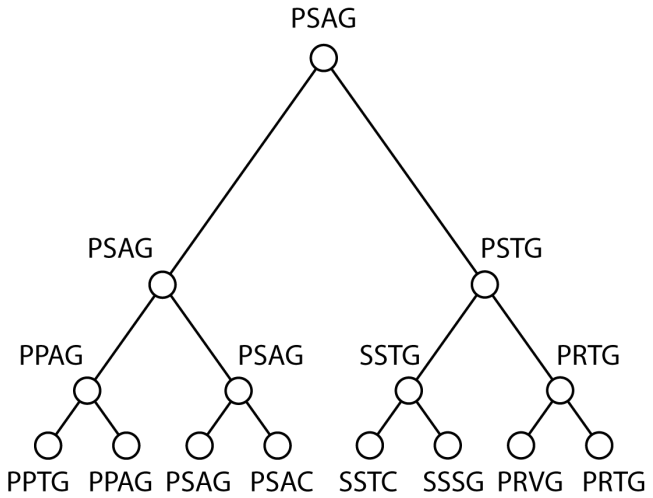
- **Joint reconstruction** – all ancestors simultaneously

- **Marginal reconstruction** – single ancestor

GDH I
GDH III
GDH II
GOx I

Why use ancestral sequence reconstruction?

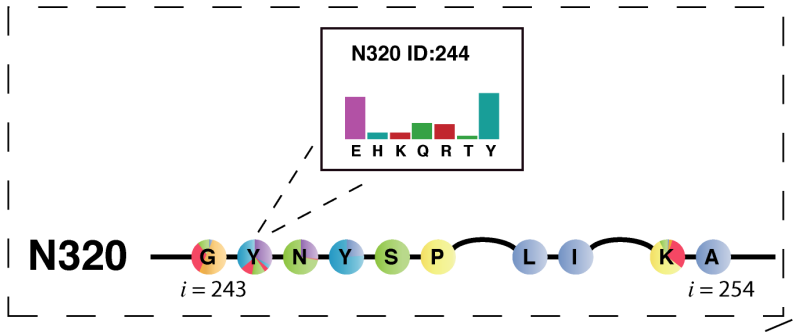
- **Studying evolutionary histories**
- Determining important functional residues
- Utilising the ancestors for industrial applications



A reconstructed history of character states

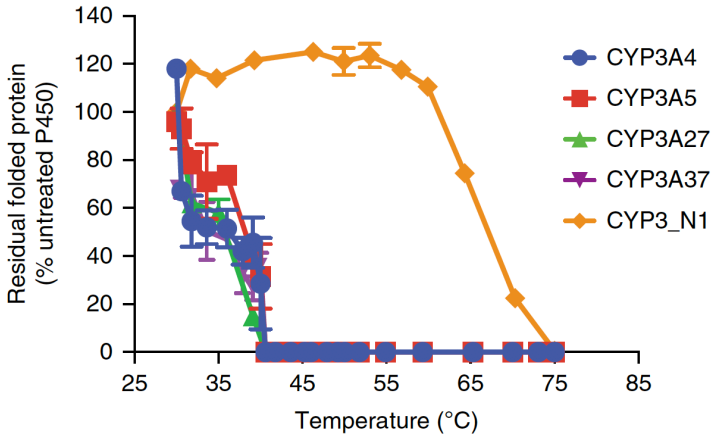
Why use ancestral sequence reconstruction?

- Studying evolutionary histories
- **Determining important functional residues**
- Utilising the ancestors for industrial applications



Why use ancestral sequence reconstruction?

- Studying evolutionary histories
- Determining important functional residues
- **Utilising the ancestors for industrial applications**



Adapted from Gumulya et al., *Nature Catalysis* **1**, 878 (2018).

Current bottlenecks

- Garcia and Kaçar (2019) reviewed 12 ASR studies from the past decade
  - Data set sizes ranged from 21 to 456 sequences
  - Average of 168 sequences

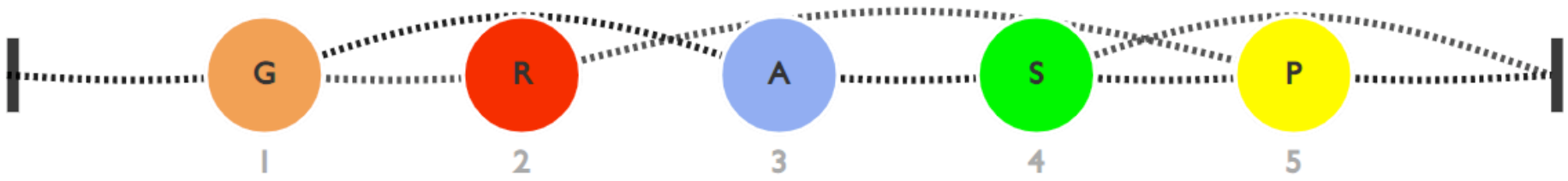
- ASR tools such as FastML, PAML are capable of ~**500-600 sequences**

- GRASP is capable of ~**10,000 sequences**

GRASP data structure and implementation

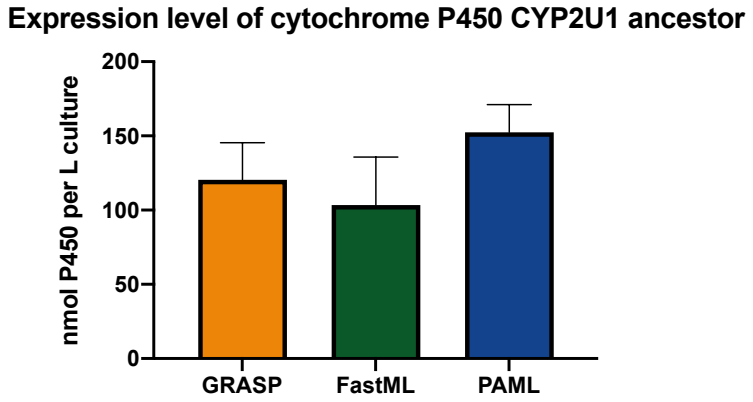
- Partial order graphs
  - Represent ambiguity
  - Summarise insertion and deletion events
- Variable elimination
  - Decompose conditional probability tables into smallest number of operations

GR--P
G-AS-
G-ASP
G-ASP



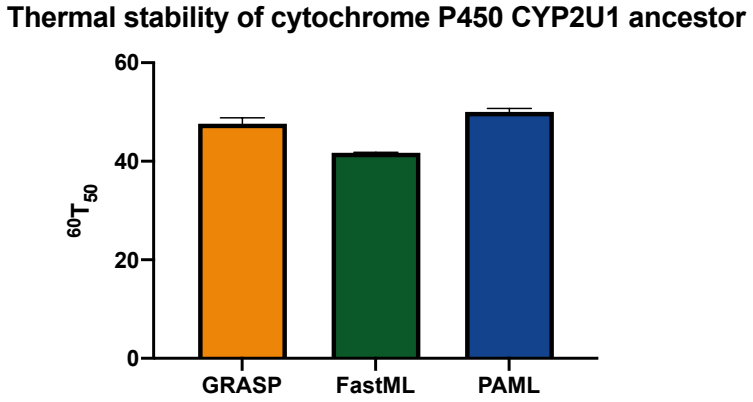
Validation – smaller data set sizes

We inferred a cytochrome P450 CYP2U1 ancestor **(359 sequences)** using GRASP, FastML, and PAML.



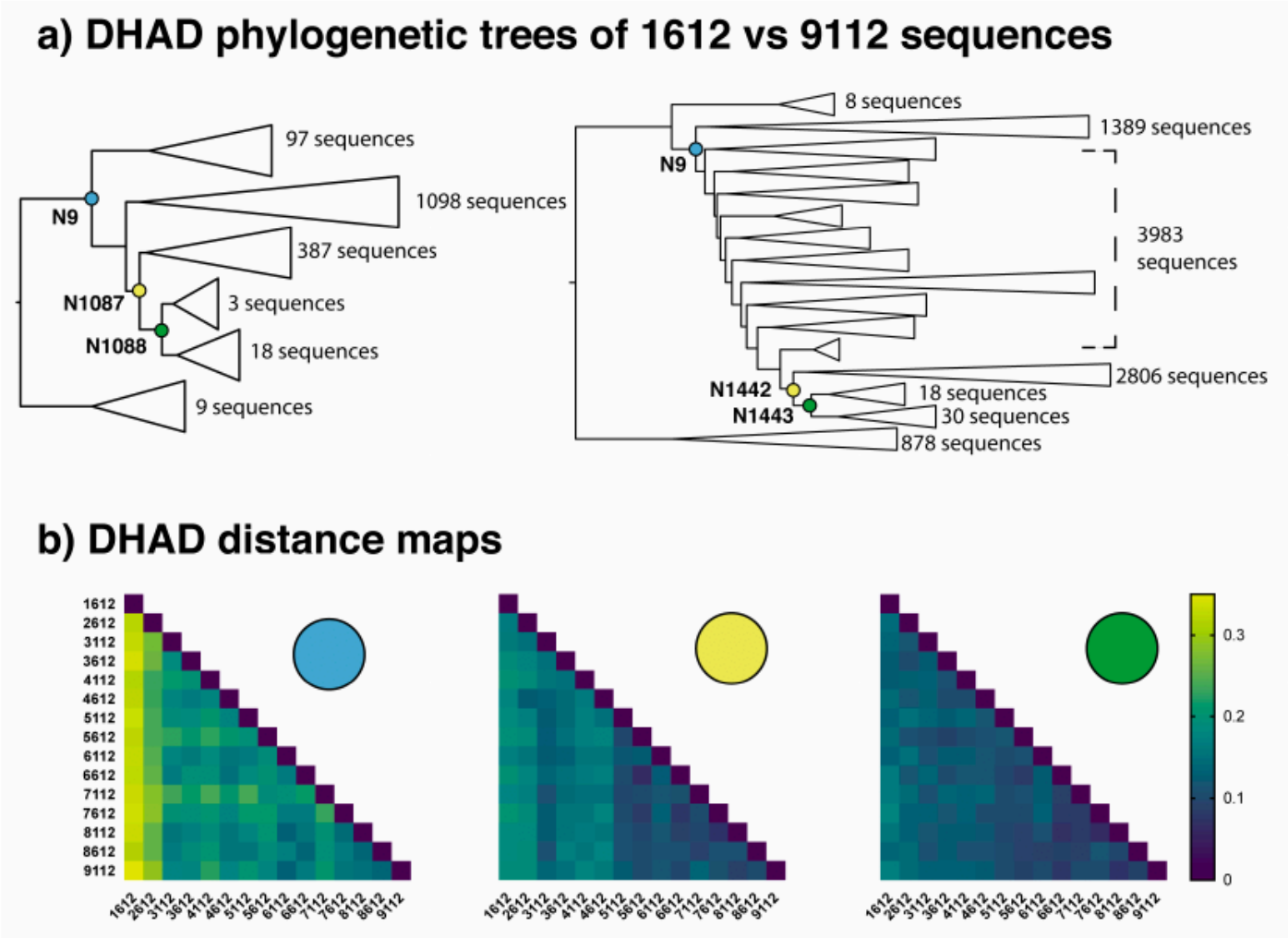
Regardless of the tool used, ancestral proteins -

- expressed at similar levels in E. coli
- displayed a P450 spectra
- had activity towards luciferin MultiCYP substrate
- showed similar thermal stabilities



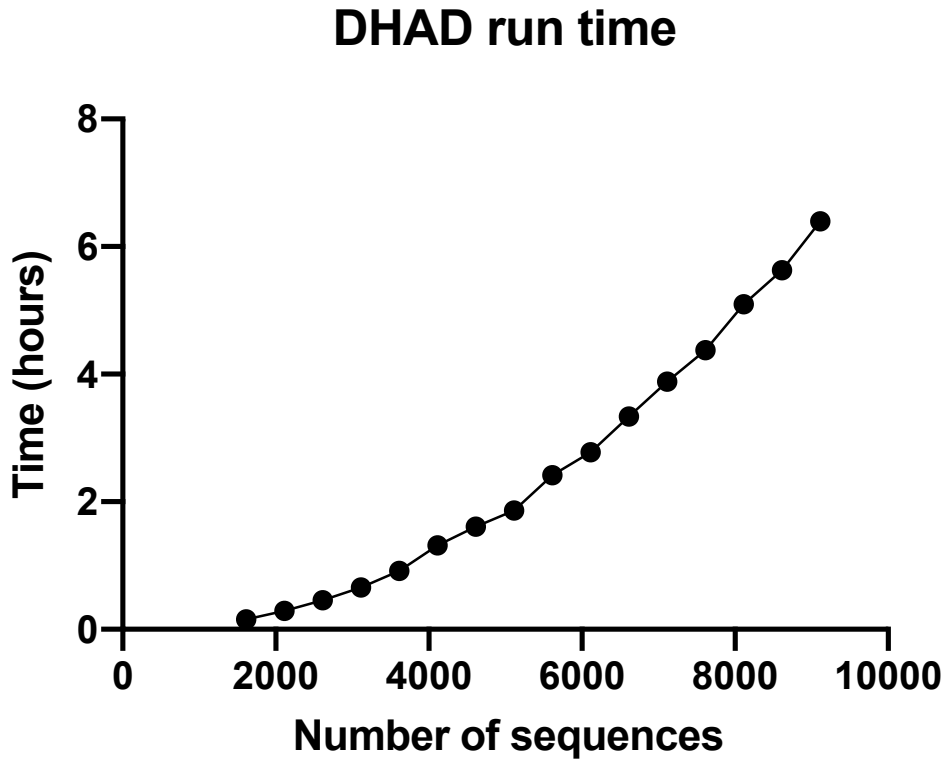
Validation – larger data set sizes

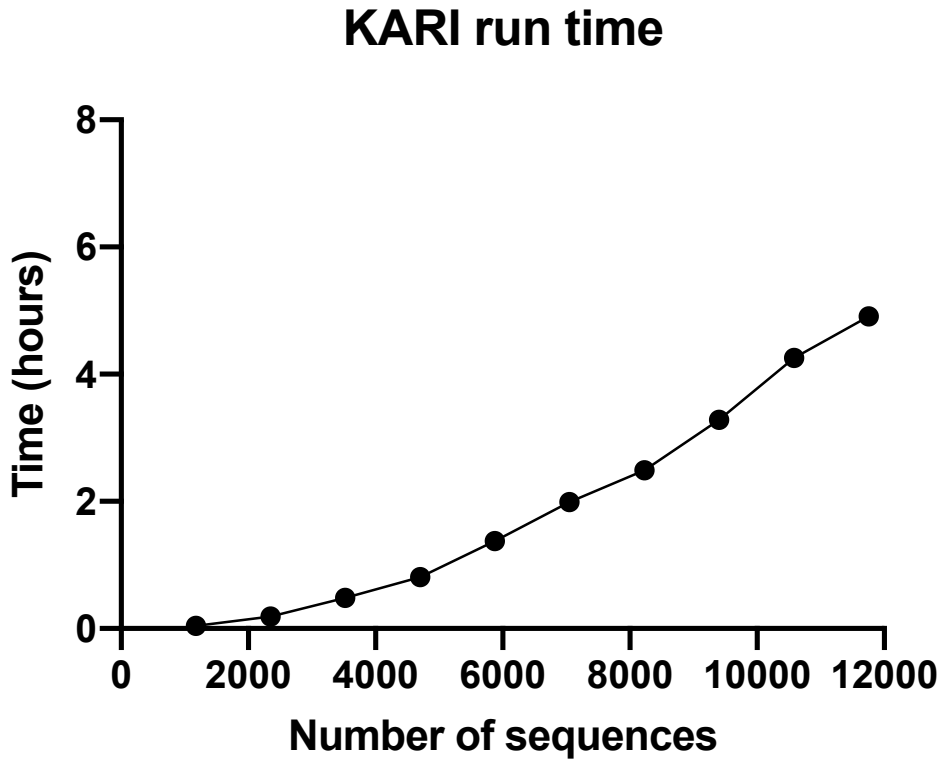
- Dihydroxy-acid dehydratase family (DHAD)
- As data set size increased ancestors were constrained towards canonical forms
- Ancestors from both the smallest and largest reconstructions were resurrected and showed activity towards D-Gluconate.



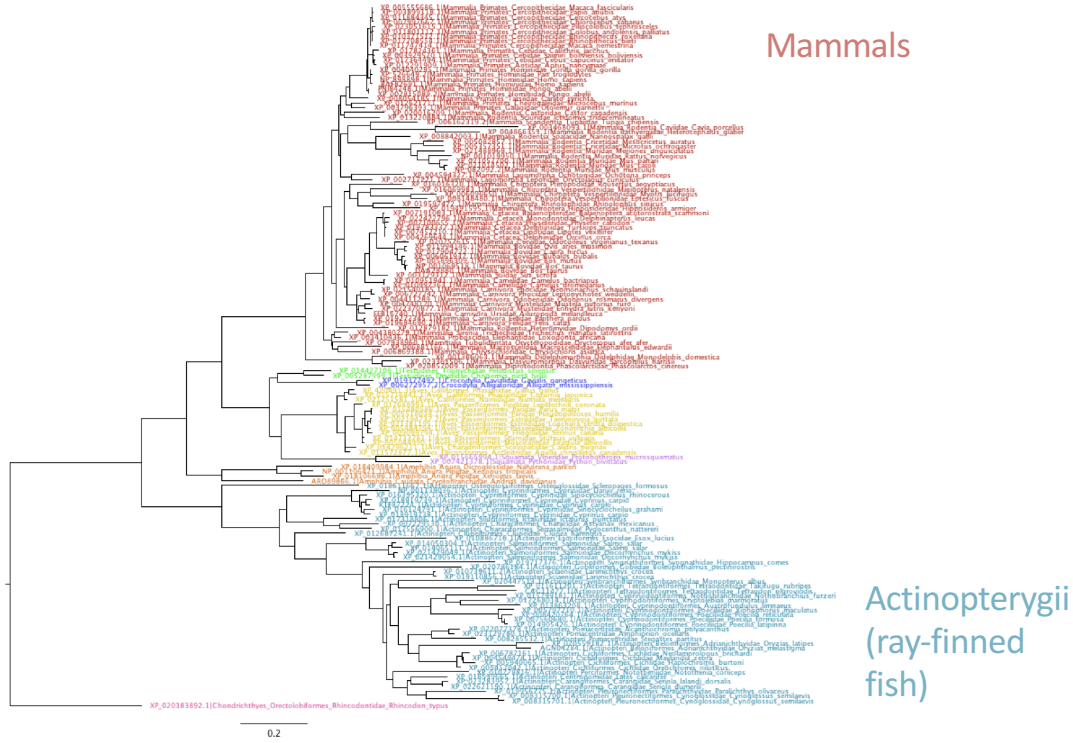
Run times

(64 GB RAM, 5 threads on 2x 2.6 GHz 14C Xeon VM)

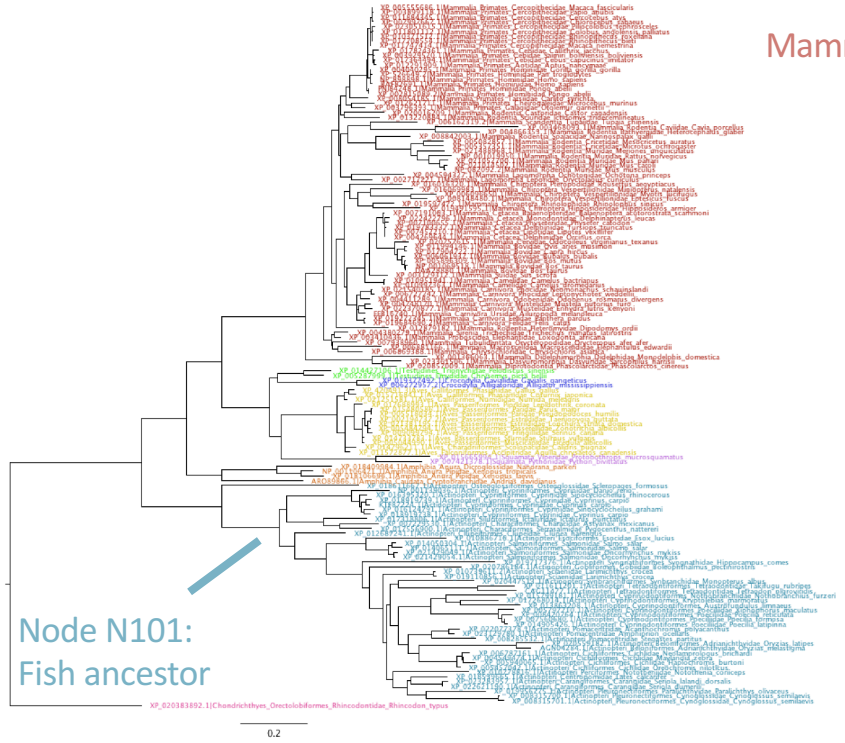




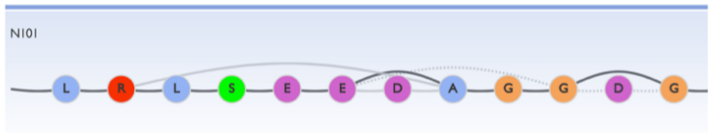
GRASP enables inspection of indel histories



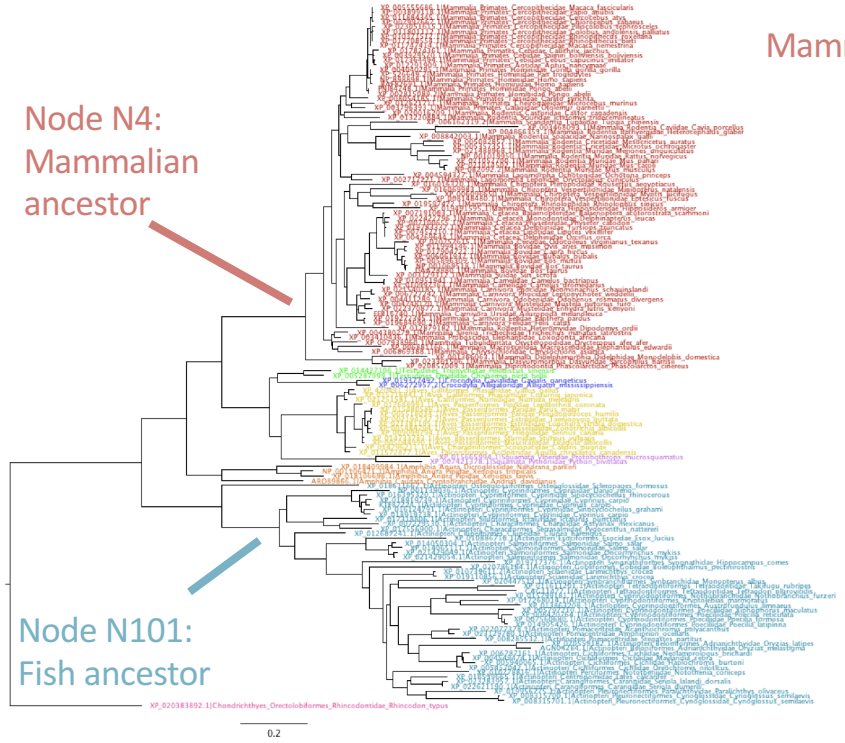
GRASP enables inspection of indel histories



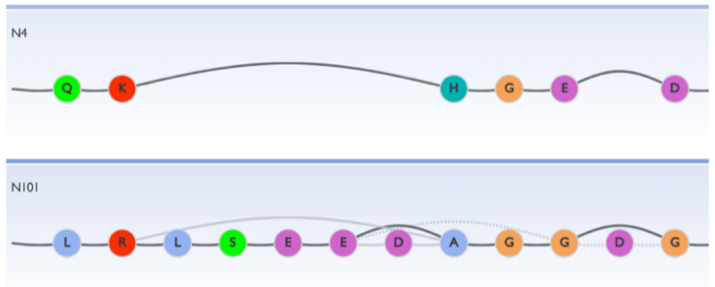
N101:



GRASP enables inspection of indel histories

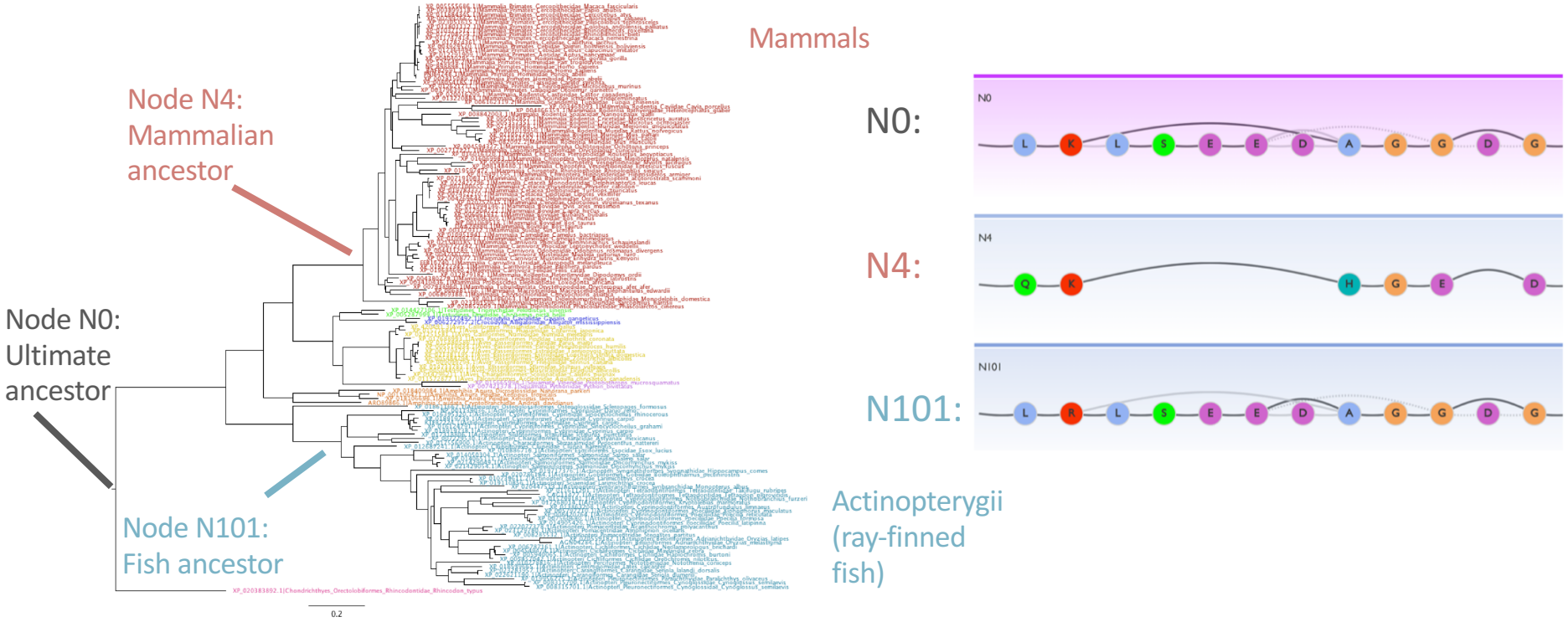


Mammals



Actinopterygii
(ray-finned
fish)

GRASP enables inspection of indel histories

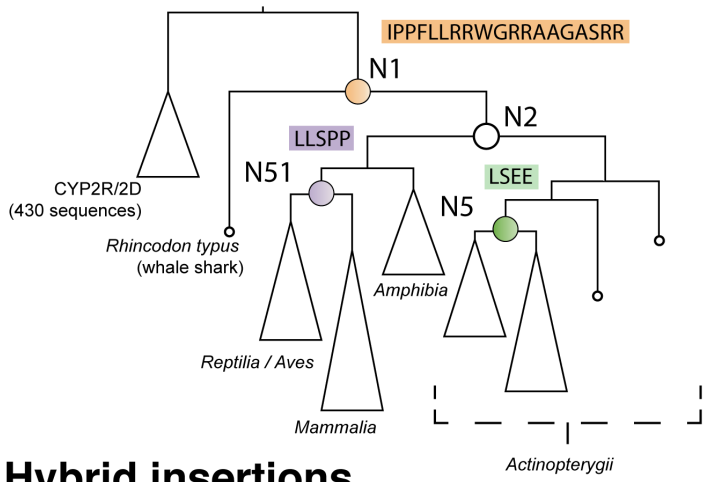


Hybrid ancestors

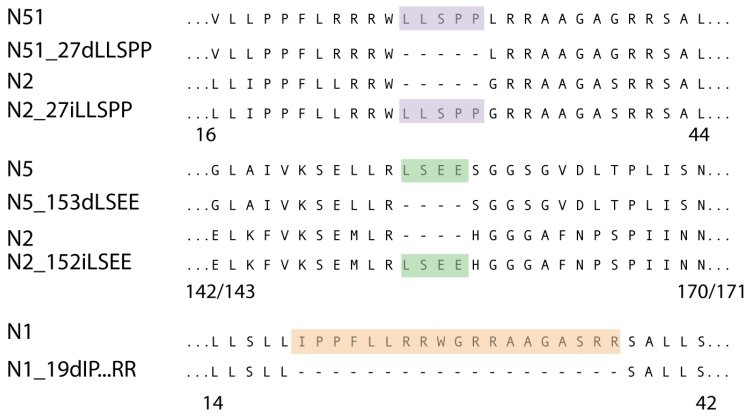
GRASP's partial order graphs allow the identification of blocks of content which can be used to create ancestral variants.

CYP2U1 variants shown to fold but with varied substrate selectivity.

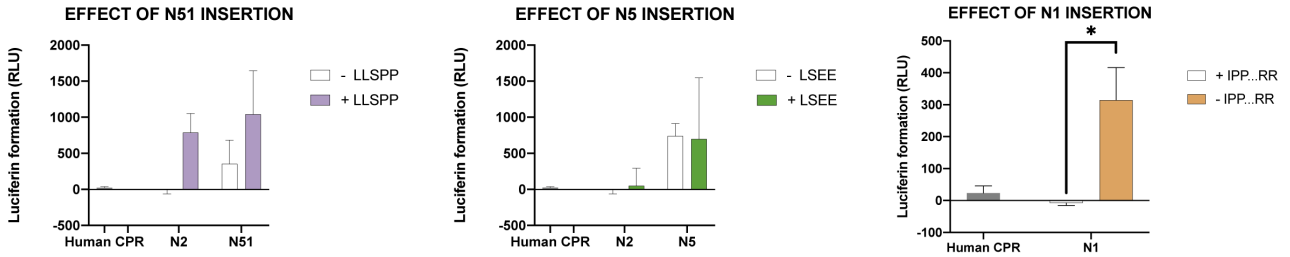
a) CYP2U/2R/2D phylogenetic tree



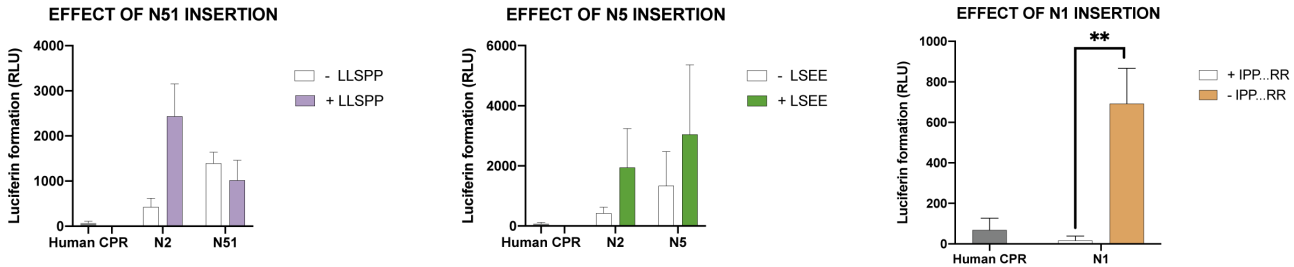
b) Hybrid insertions



c) Activity with luciferin CEE

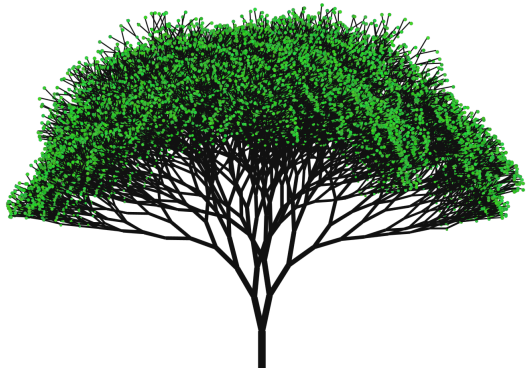


d) Activity with luciferin ME-EGE



Conclusion

- Ancestral sequence reconstruction is a valuable resource to understand, explore, and utilise evolution
- Large data sets allow us to extend the reach of ASR
- GRASP enables novel experiments on previously unobtainable data set sizes



Acknowledgements

GRASP implementation

Mikael Bodén

Ariane Mora

Marnie Lamprecht

Julian Zaugg

Alexandra Essebier

Brad Balderson

Rhys Newell

CYP2U1 experimental work

Elizabeth Gillam

Connie Ross

Raine Thomson

Ross Barnard

Luke Guddat

Gary Schenk

Bostjan Kobe

Burkhard Rost

DHAD experimental work

Volker Sieber

Scott Bottoms

Jörg Carsten

GMC experimental work

Dietmar Haltrich

Leander Sützl

GRASP

Graphical representation of ancestral sequence predictions

http://grasp.scmb.uq.edu.au

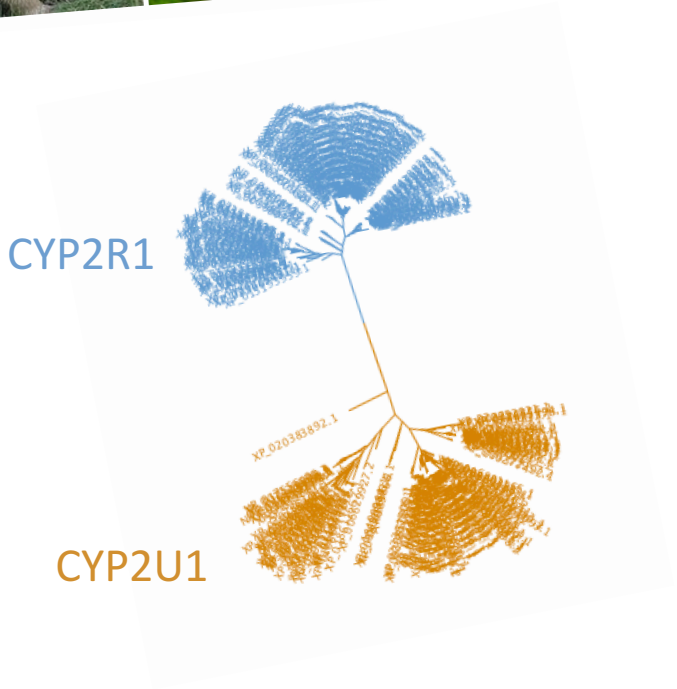
SCAN ME

Additional slides

Cytochrome P450 2U1 subfamily

Cytochrome P450 enzymes are members of a superfamily of monooxygenases that play a **critical role in metabolism**

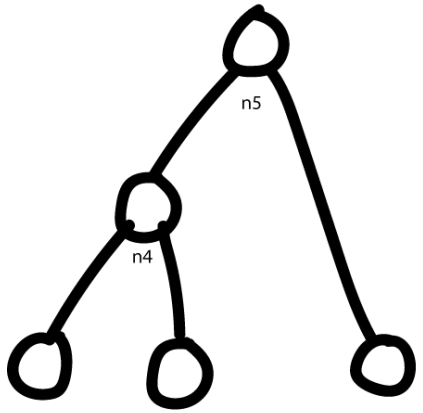
CYP2U1 – cytochrome P450 subfamily found across, amphibians, reptiles, mammals, birds, and fish.



CYP2U1 is interesting because –

- No exact established function and substrate specificity
- Previous cytochrome P450 ancestors showed increased stability and promiscuity

Marginal & joint differences



P(n4 = A, n5 = A) = 0.4
P(n4 = A, n5 = C) = 0.3
P(n4 = C, n5 = A) = 0.05
P(n4 = C, n5 = C) = 0.25

Joint reconstruction of node n4 and node n5
Find the highest probability
P(n4 = A, n5 = A) = 0.4

Character at n5 is assigned A

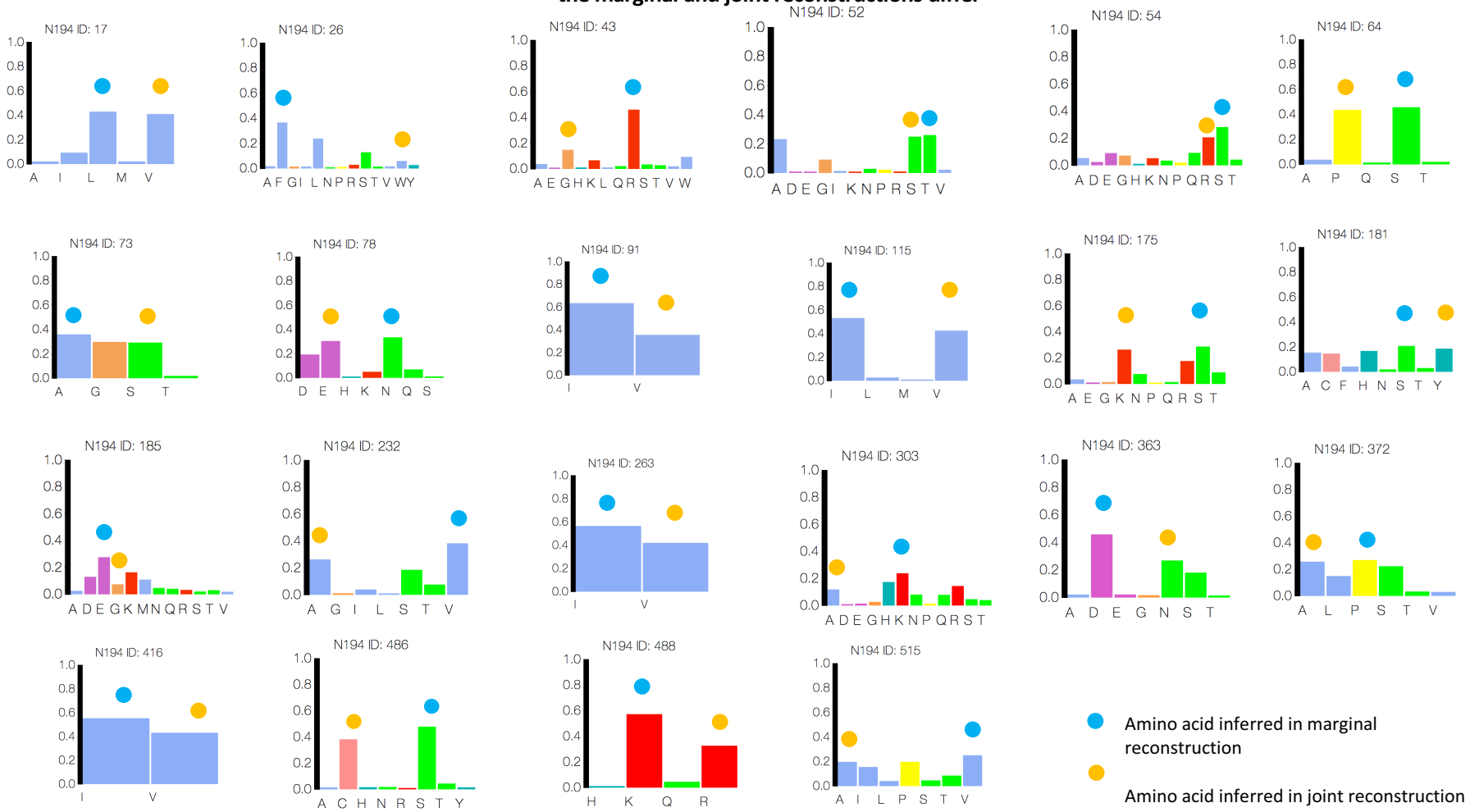
Marginal reconstruction of node n5
Sum up all the ways we could get n5=A
P(n4 = A, n5 = A) + P(n4= C, n5 =A)
= 0.4 + 0.05
= 0.45

Sum up all the ways we could get n5=C
P(n4 = A, n5 = C) + P(n4 = C, n5 = C)
= 0.3 + 0.25
= 0.55

Character at n5 is assigned C

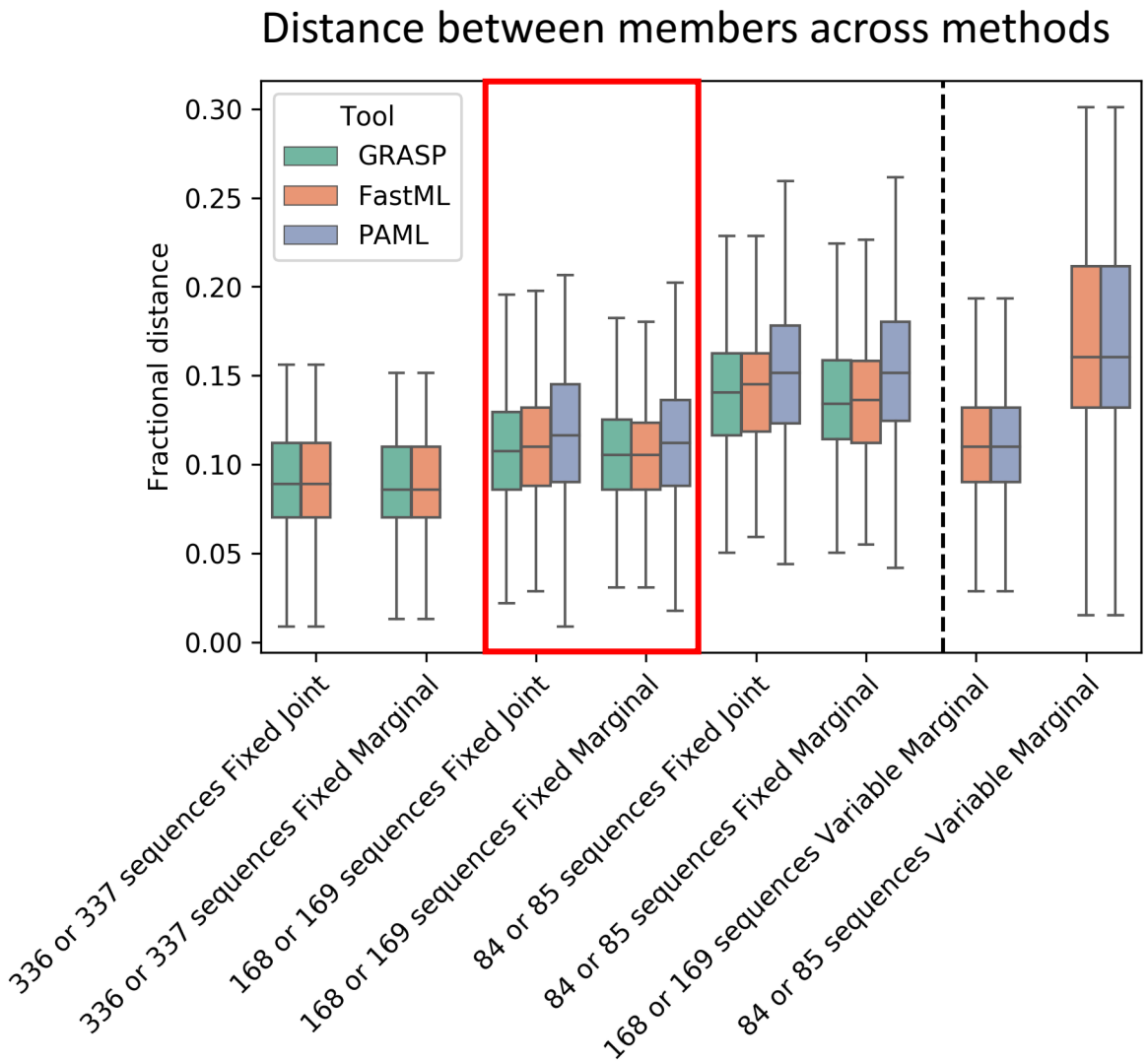
Marginal & joint differences

Posterior probability distributions from the CYP2U1 CYP2R1 Realigned marginal reconstruction at positions where the marginal and joint reconstructions differ



Methods' consensus standard:

Generate ancestors that are similar to those of other methods



Data consensus standard:

Generate ancestors close to that of the superset