Towards an integrative approach to the investigation of linguistic phylogenies

Munich-UCLA Historical Linguistics Colloquium V

M.I. Rehan

2024-06-28

Introduction

- In this presentation, I want to ask (and try to answer) two related questions.

- In this presentation, I want to ask (and try to answer) two related questions.
  - Do different components of language evolve at a statistically-quantifiable homogeneous or heterogeneous rate?
  - How can we best integrate all the linguistic evidence (e.g., morphological, syntactic, phonological, lexical) available to us for phylogenetic inference?
- While the questions are simply stated, the answers are anything but trivial and rarely investigated.

- I attempt to answer these questions by running various Bayesian phylogenetic models on the dataset of ancient Indo-European languages created by Nakhleh et al. (n.d.), which has the advantage of being one of the only datasets with lexical, phonological, and morphological characters.
- In my analyis, I will use a Bayesian modeling framework known as a **Partitioned/Mixed-Model Approach** popularized by Brandley et al. (2005).
    - Denison et al. (2002) provides a review of the partitioned Bayesian models.

- Computational Phylogenetics

- Computational Phylogenetics
- The Data

- Computational Phylogenetics
- The Data
- The Partitioned Models

- Computational Phylogenetics
- The Data
- The Partitioned Models
- Results and Discussion

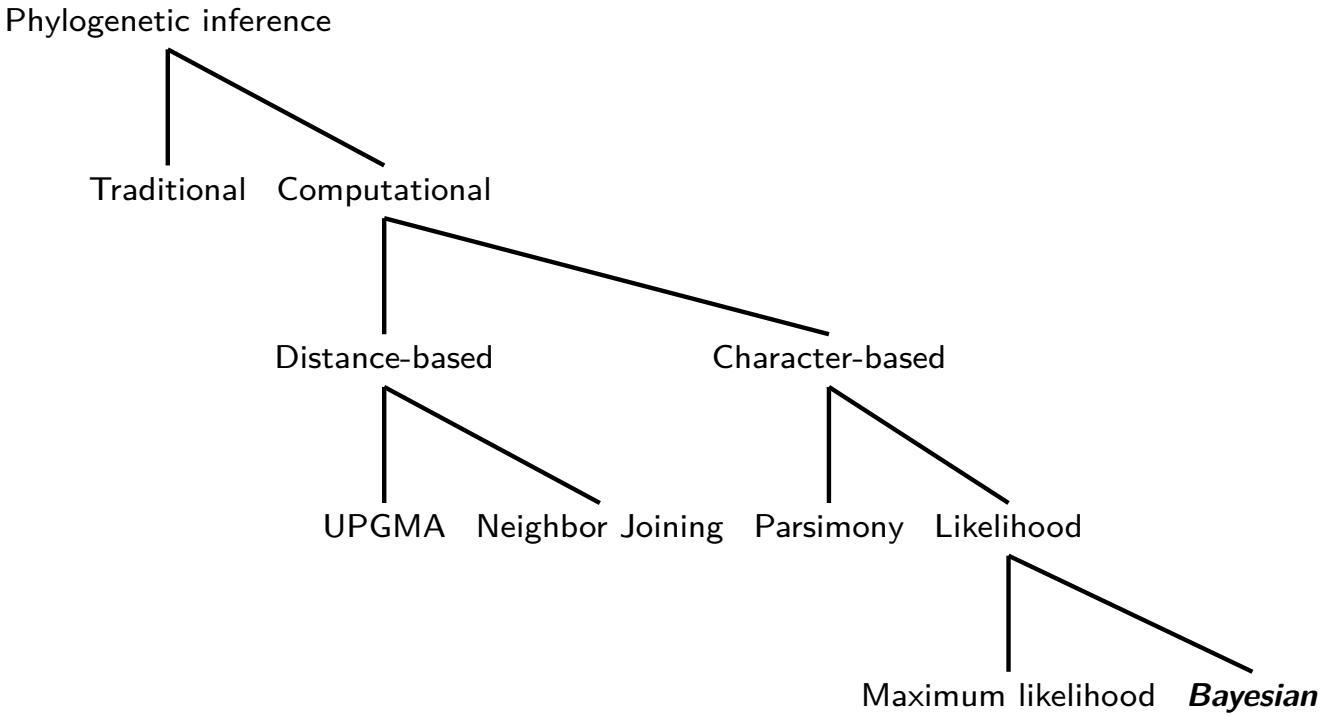
- Computational Phylogenetics
- The Data
- The Partitioned Models
- Results and Discussion
- Conclusions

Computational Phylogenetics

- The turn of the 21st century saw the development of a new research kit for the investigation of the Indo-European family tree: **computational cladistics/phylogenetics**.

- The turn of the 21st century saw the development of a new research kit for the investigation of the Indo-European family tree: **computational cladistics/phylogenetics**.
- Some early computational research into Indo-European phylogeny was completed by the research group of Tandy Warnow and Don Ringe (Warnow 1997, Ringe et al. 1998, Ringe 2000, Ringe 2002, Ringe et al. 2002) and continued by them and successive iterations of colleagues resulting in an impressive collection of papers from their Computational Phylogenetics in Historical Linguistics (CPHL) group (Warnow n.d.).

- The turn of the 21st century saw the development of a new research kit for the investigation of the Indo-European family tree: **computational cladistics/phylogenetics**.
- Some early computational research into Indo-European phylogeny was completed by the research group of Tandy Warnow and Don Ringe (Warnow 1997, Ringe et al. 1998, Ringe 2000, Ringe 2002, Ringe et al. 2002) and continued by them and successive iterations of colleagues resulting in an impressive collection of papers from their Computational Phylogenetics in Historical Linguistics (CPHL) group (Warnow n.d.).
- These computational phylogenetic analyses can be broadly divided into two categories: character-based, distance-based.



- For comparisons of these different methods, see Barbancon et al. (2013), Canby et al. (2024), Goldstein (2020)
- For the use of another modeling approach called *Agent-Based Modeling* for phylogenetic inference, see Hartmann 2023, Sandell 2023, Hartmann Forth.

- Gray and Atkinson (2003) is the first investigation of Indo-European phylogeny in a Bayesian framework.
  - Their analysis supported the Antolian hypothesis of PIE origin.
  - Chang et al. (2015) on the other hand, supports the Steppe hypothesis of Indo-European origin by a revised version of the IE-LEX dataset (available at IELEX 2023) and constraints on ancestry.

- Heggarty et al. (2023) support a hybrid hypothesis of Indo-European origin by using the newly published IE-CoR dataset (Heggarty et al. 2024) The major difference is the sampling of language ancestry by the use of a **Fossilized-Birth-Death Model** (FBD) of ancestry (for the use of an FBD model for Romance DTE, see Goldstein 2024 with references and Rama 2018).
    - No ancient language is constrained to be the ancestor of a modern language. This allows the model to treat an extinct sister of an ancient language as the direct ancestor of a modern language.

- These analyses rely solely on one type of linguistic data: cognacy relationships, which are problematic and have been problematized in the past (cf. most recently Abner et al. (2024)).

- Little attention has been paid to the incorporation of phonological, morphological, and syntactic characters alongside cognacy relationships for the investigation of phylogeny, except in the work of Ringe and colleagues (e.g., Nakhleh et al. (2005)), whose curated dataset I will be using for my own analysis.

The Data

- An example of a morphological character:

Hitt.	2	**Av.**	1	**Luv.**	10	**Goth.**	15
Arm.	1	**OCS**	5	**Lyc.**	11	**ON**	16
Gk.	1	**Lith.**	6	**TA**	12	**OHG**	17
Alb.	3	**OE**	7	**OPer.**	13	**Welsh**	18
TB	4	**OI**	8	**OPru.**	13	**Osc.**	19
Ved.	1	**Lat.**	9	**Latv.**	14	**Umb.**	20

Table 1: M2 augment: 1 = present, 2 and cont. = absent

- Every distinct number marks a character state.

Hitt.	2?	**Av.**	1	**Luv.**	1	**Goth.**	1
Arm.	2?	**OCS**	1	**Lyc.**	4	**ON**	1
Gk.	1	**Lith.**	1	**TA**	1	**OHG**	1
Alb.	3	**OE**	1	**OPer.**	1	**Welsh**	5
TB	1	**OI**	2	**OPru.**	1	**Osc.**	2
Ved.	1	**Lat.**	2	**Latv.**	1	**Umb.**	2

Table 2: M11 abstract noun suffix [Ringe et al. 2003 SM: 9]

Table 3: P2 full "satem" development (labiovelars merge with velars; PIE "palatals" become affricates or fricatives)

Hitt.	1	Av.	2	Luv.	1	Goth.	1
Arm.	1	OCS	2	Lyc.	1	ON	1
Gk.	1	Lith.	2	TA	1	OHG	1
Alb.	1	OE	1	OPer.	2	Welsh	1
TB	1	OI	1	OPru.	2	Osc.	1
Ved.	2	Lat.	1	Latv.	2	Umb.	1

- The real evidence is more problematic: for example, Anatolian languages can still show conditioned outcomes of the palatals (Melchert 2012) despite the fact that they are supposed to have been merged with the velar series.

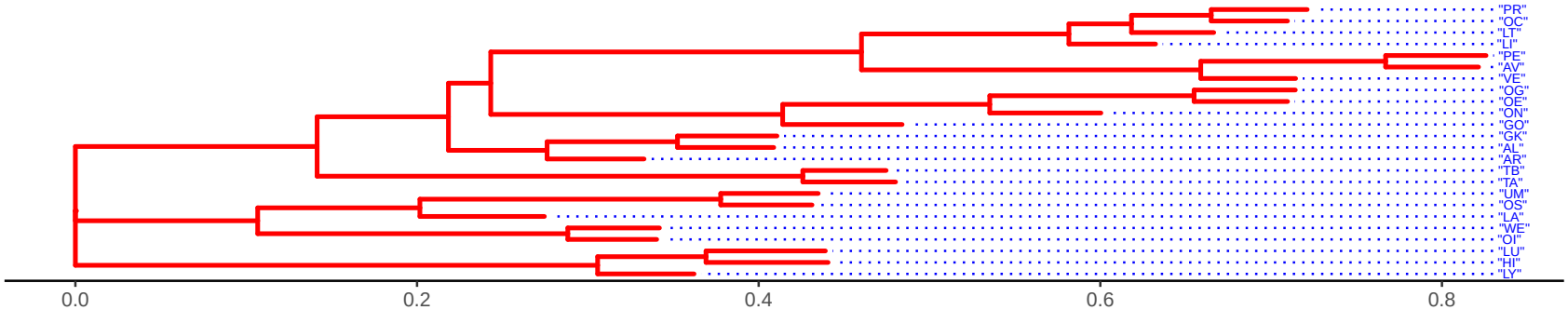
- The real evidence is more problematic: for example, Anatolian languages can still show conditioned outcomes of the palatals (Melchert 2012) despite the fact that they are supposed to have been merged with the velar series.
- Indic can also show conditioned reflexes of the labiovelar series, which as a satem-language, it is supposed to have merged with the velars: *$g^{w}_{\circ}rh_2ú$-* > *gurú-* 'heavy' (cf. Gk. βαρύς 'heavy', Lat. *gravis* 'id'. For a recent treatment of the outcomes of labiovelars in Vedic, see (Clayton 2022).

- These multi-state character sets are converted in the binary format using the concept of cognate classes to be used for further inference.
- There are methods out there to use multi-state data (Canby 2024, some idiosyncratic Heggarty et al. 2023), but I am afraid that this dataset might not be conducive to those methodologies.

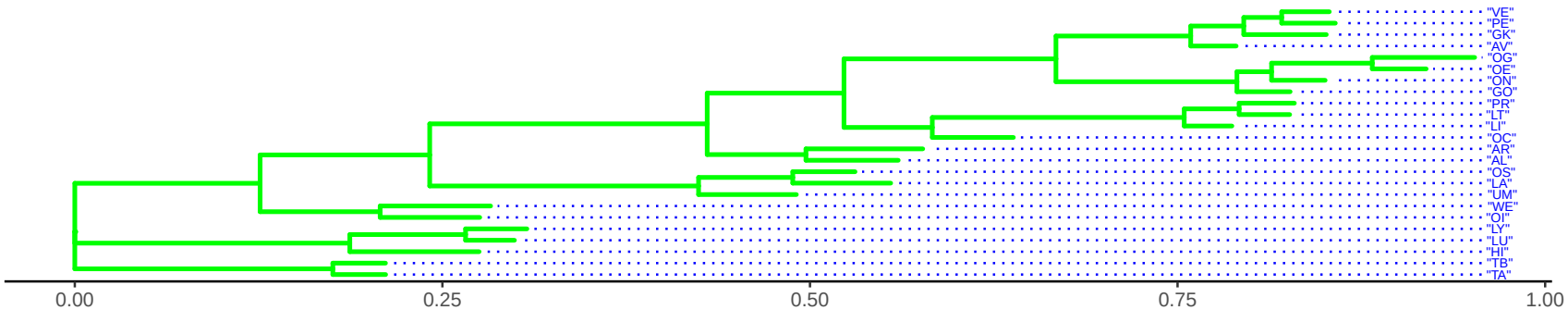
- I ran a preliminary analysis[1] on these separate sets of linguistic characters, and the phylogeny inferred by morphological characters is at odds with the phylogeny inferred from phonological and lexical characters.
- 'Some degree of difference is not unexpected: "these different levels of language need not necessarily evolve in tandem and remain fully aligned." Heggarty et al. (2023: 66-67)

[1] F-81+G+I

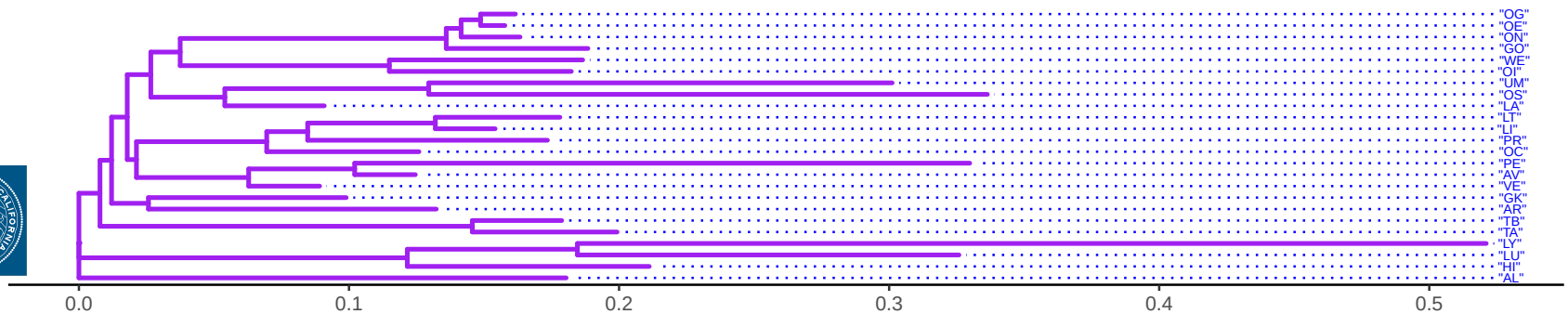
Tree Inferred from Phonological Characters



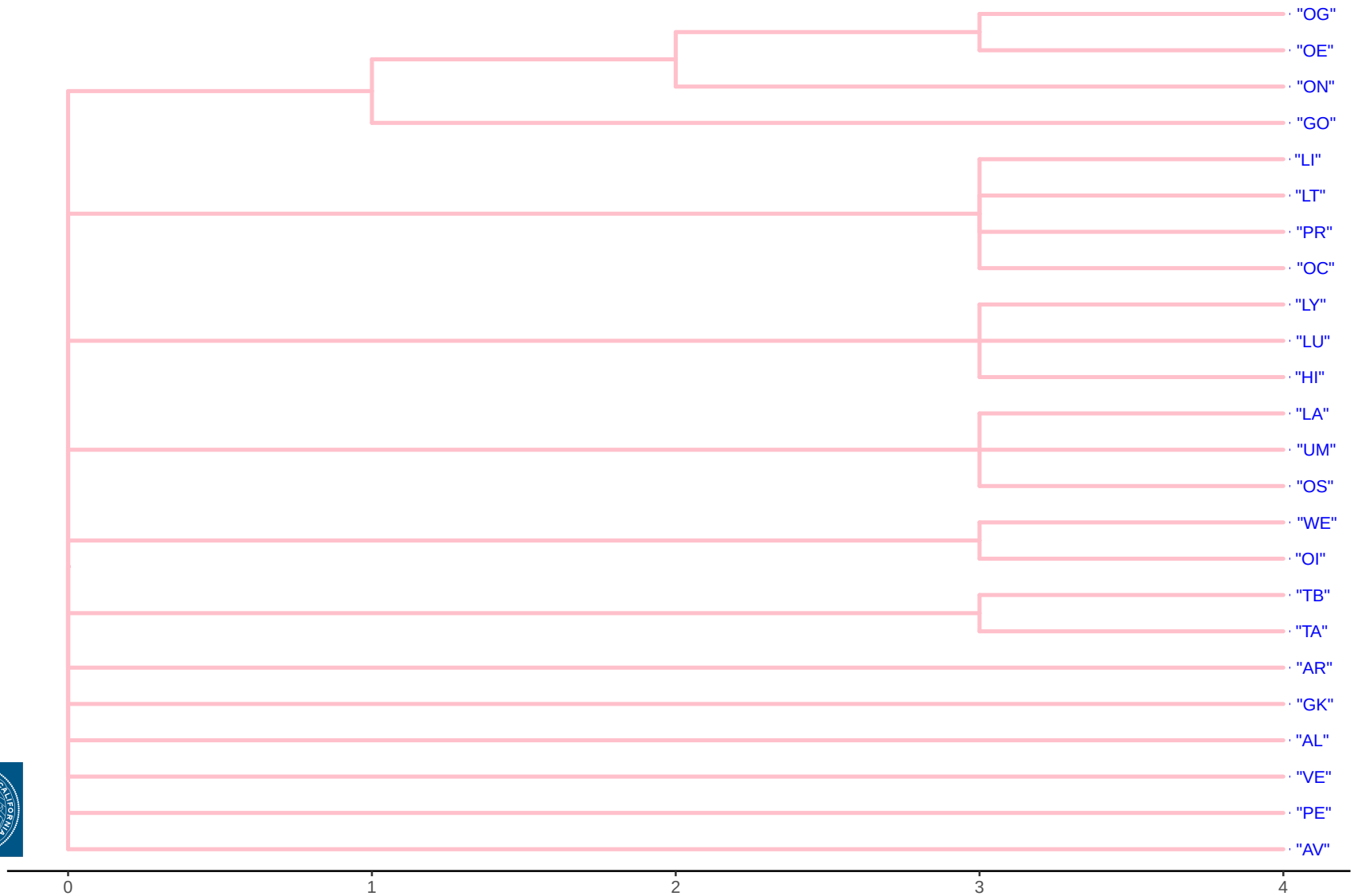
Tree Inferred from Morphological Characters



Tree Inferred from Lexical Characters



Consensus Tree



Partitioned Models

The modeling approach used so far assumes that all categories of linguistic characters evolve under the same evolutionary dynamics.

- However, Nakhleh et al. (2005) contend that morphological characters are more resistant to change and, in their words, "there is ample evidence that these [sc. morphological] characters coded here are far likelier to reflect the true tree."

Instead of assuming the same evolutionary dynamics for phonological, morphological, and lexical characters, we can split the problem of inference into parts by modeling the evolution of each category of linguistic character separately and then jointly inferring a tree based on all characters.

We partition the data into alignments that we believe to evolve under the same evolutionary processes -> separate alignments for different linguistic characters.

We partition the data into alignments that we believe to evolve under the same evolutionary processes -> separate alignments for different linguistic characters.

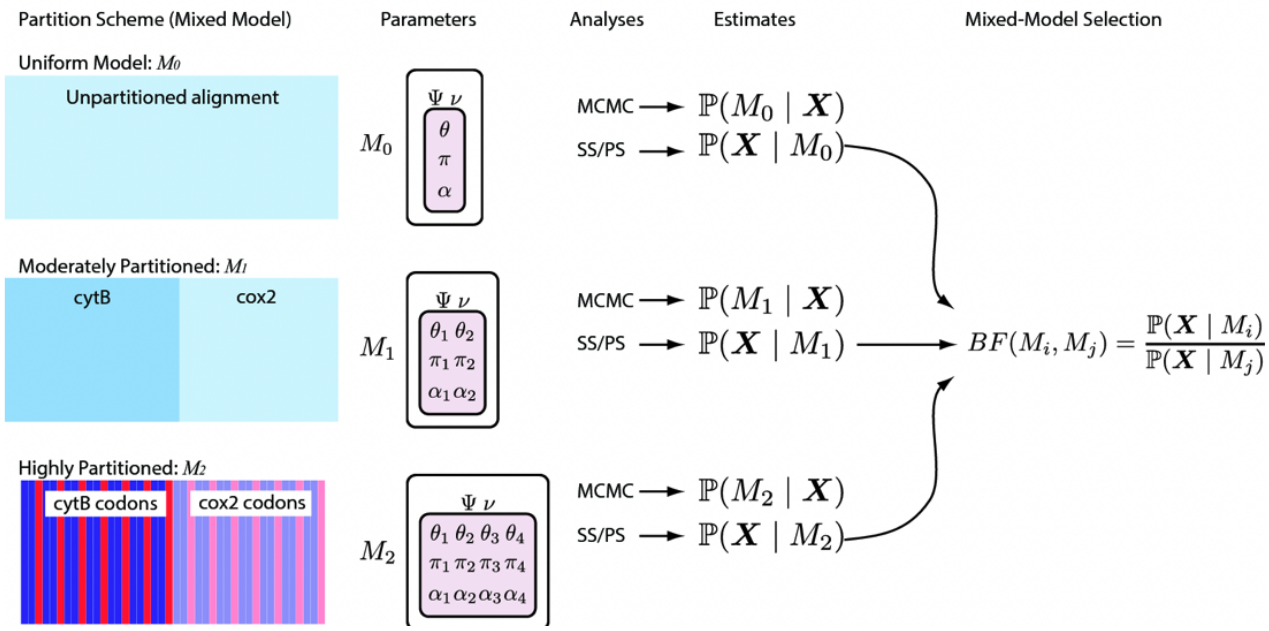


Figure 1: An overview of the partitioned model

We model the evolution of character states as a **Continuous-Time Markov Chain**.

- Across the partitioned and unpartitioned analyses, I emphasize three parts of the model:
  - An Instantaneous transition rate matrix $Q_{ij} = Q_{ji}$
  - A model of stationary frequencies
  - A model of rate variation among (discrete group of) sites

For the simplest model, we get the following Q (instantaneous-rate) matrix:

$$Q = \begin{pmatrix} -\alpha & \alpha \\ \alpha & -\alpha \end{pmatrix}$$

- The model that I chose in the end for my partitioned analysis has gamma-distributed sites (4 discrete partitions), equal bidirectional transition probabilities (i.e., $Q_{(ij)} = Q_{(ji)}$), and the stationery frequencies are estimated from the data.

- The model that I chose in the end for my partitioned analysis has gamma-distributed sites (4 discrete partitions), equal bidirectional transition probabilities (i.e., $Q_{(ij)} = Q_{(ji)}$), and the stationery frequencies are estimated from the data.

$$Q_k = \begin{pmatrix} -\alpha\pi_1 & \alpha\pi_1 \\ \alpha\pi_0 & -\alpha\pi_0 \end{pmatrix}$$

- In this model, π_0 & π_1 are estimated from the data and not assumed to be equal.

- The model that I chose in the end for my partitioned analysis has gamma-distributed sites (4 discrete partitions), equal bidirectional transition probabilities (i.e., $Q_{(ij)} = Q_{(ji)}$), and the stationery frequencies are estimated from the data.

$$Q_k = \begin{pmatrix} -\alpha\pi_1 & \alpha\pi_1 \\ \alpha\pi_0 & -\alpha\pi_0 \end{pmatrix}$$

- In this model, π_0 & π_1 are estimated from the data and not assumed to be equal.

- But what if we further think that the transition rate from a character state A -> character state B does not happen at the same rate as the change from B -> A, then we can incorporate this asymmetry into our model by using a generalized-time reversible model:

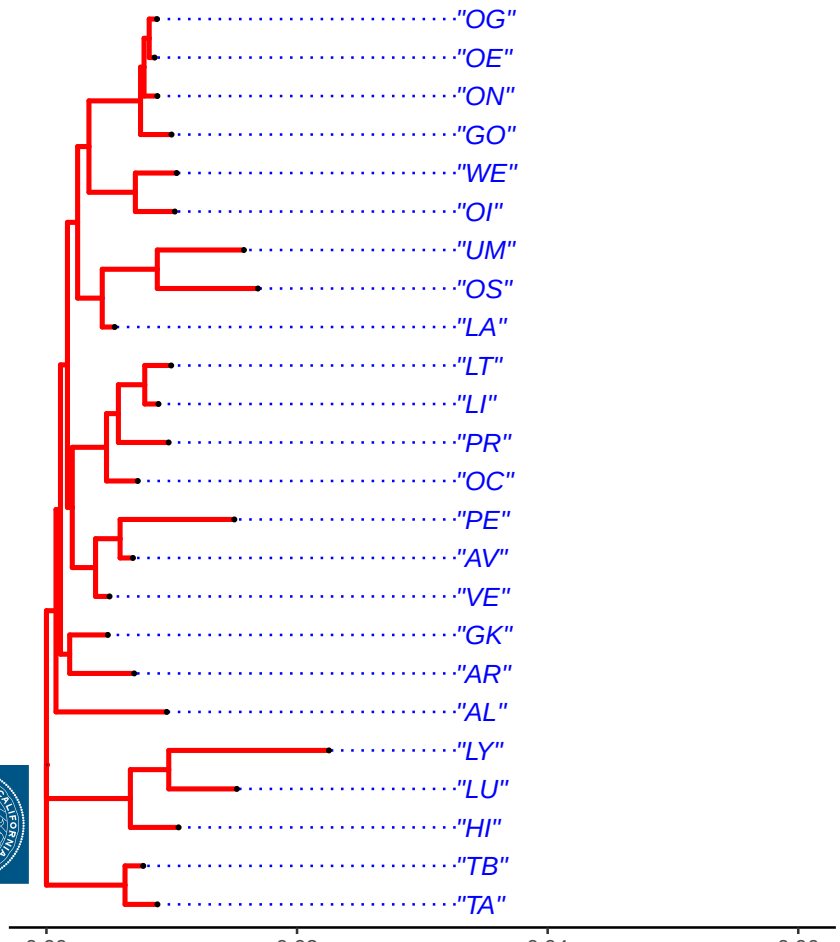
- But what if we further think that the transition rate from a character state A -> character state B does not happen at the same rate as the change from B -> A, then we can incorporate this asymmetry into our model by using a generalized-time reversible model:
  - A central assumption of this model is the time-reversibility: $Q_{ij}\pi_i = Q_{ji}\pi_j$

$$Q_k = \begin{pmatrix} -\alpha\gamma_k\pi_1 & \alpha\gamma_k\pi_1 \\ \alpha\gamma_k\pi_0 & -\alpha\gamma_k\pi_0 \end{pmatrix}$$

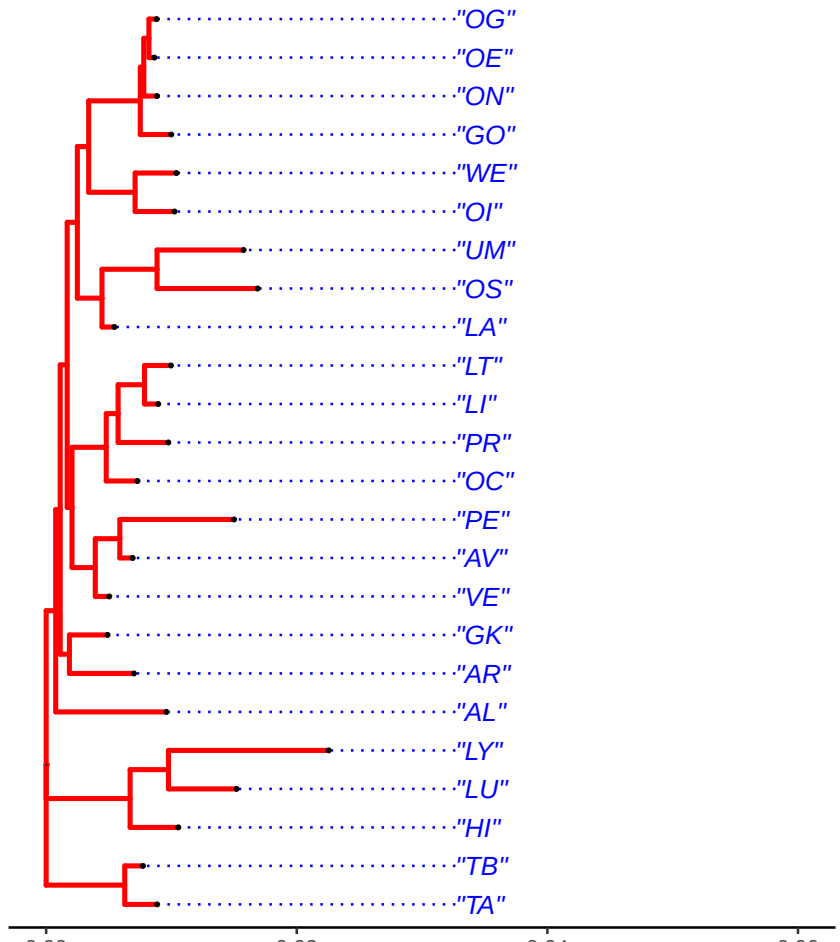
- We simply scale the transition probability of a given character by the gamma-rate category γ_k to which the site belongs.

Results and Discussion

MCC Tree



MAP Tree



- We recover all the major clades and a few higher level subgroups.
- Fragmentary languages, however, have disproportional branch lengths: Lycian, Oscan, Umbrian, Old Persian.
  - This most likely reflects an artifact of the model estimating more change because of ? sites not being informative.

- Since my trees are ultrametric and are not time-calibrated, it is not possible to compare the results of the model in terms of their correspondence to the agreed-upon sub-groupings established by traditional comparative methods.
  - For model comparison, we will be using **Bayes Factor** which is calculated as the ratio of the marginal likelihoods of the models under comparison.
  - Simply stated, Bayes Factor is a measure of the *relative fit* of the model to the data.
    - Hence, it makes no claim about the *adequacy* of the model itself.

Unpartitioned	**Marginal Likelihoods**
JC	-21920.26
F-81 V-Code	-22084.31
GTR+G+I	**-19352.41**
GTR+G+I+GBI	-19354.12
Partitioned	**Marginal Likelihoods**
F81+ASRV(4)	**-19110**
JC	-21989.58
Kappa-JC-JC	-21997.82

$$\mathrm{BF}_{\mathsf{F81+ASRV(4),GTR+G+I}} = 242.41$$

Conclusions

- My analysis of the Ringe dataset provides evidence in favor of a partitioned F-81+ASRV(4) model over a suite of unpartitioned analyses, which suggests that an integrative approach might be possible way.
    - Some morphological characters change at the same rate as phonological characters while others do not, so we also have a need for better alignments.

- My analysis of the Ringe dataset provides evidence in favor of a partitioned F-81+ASRV(4) model over a suite of unpartitioned analyses, which suggests that an integrative approach might be possible way.
  - Some morphological characters change at the same rate as phonological characters while others do not, so we also have a need for better alignments.
    - While biologists have developed heuristics to investigate the optimal partition alignments (Lanfear et al. 2014) the use of partitioned alignments needs to be thoroughly investigated in linguistic phylogenetics.

- For the analysis to hold water, it needs to replicated across a multitude of datasets, but here's the rub:
  - There is a dearth of datasets made by informed Indo-Europeanists like Don Ringe and, hence, we are in a situation where most of the Bayesian phylogenetics research does not incorporate all of the evidence available to us in phylogenetic inference.
  - I hope this presentation might convince some to not slight morphological and phonological characters for phylogenetic inference.

Thanks for Your Attention!

2

[2] I would like to thank David Goldstein and John Clayton for their help with various aspects of this presentation.

Abner, Natasha, Grégoire Clarté, Carlo Geraci, Robin J. Ryder, Justine Mertz, Anah Salgat, and Shi Yu. 2024. "Computational Phylogenetics Reveal Histories of Sign Languages." *Science* 383 (6682): 519–23. https://doi.org/10.1126/science.add7766.

Clayton, John. 2022. "Labiovelar Loss and the Rounding of Syllabic Liquids in Indo-Iranian." *Indo-European Linguistics* 10: 33–87. https://brill.com/ieul.

Nakhleh, L., T. Warnow, D. Ringe, and S. N. Evans. 2005. "A Comparison of Phylogenetic Reconstruction Methods on an IE Dataset." *Transactions of the Philological Society* 3 (2): 171–92.