

# Computational Approaches for Integrating out Subjectivity in Cognate Synonym Selection

**Luise Häuser**

Computational Molecular Evolution group,  
Heidelberg Institute for Theoretical  
Studies, Heidelberg, Germany,  
Institute for Theoretical Informatics,  
Karlsruhe Institute of Technology, Karlsruhe, Germany  
luise.haeuser@h-its.org

**Gerhard Jäger**

University of Tübingen  
gerhard.jaeger@uni-tuebingen.de

**Alexandros Stamatakis**

Biodiversity Computing Group,  
Institute of Computer Science, Foundation for Research and Technology - Hellas  
Computational Molecular Evolution group,  
Heidelberg Institute for Theoretical Studies, Heidelberg, Germany Institute for Theoretical Informatics,  
Karlsruhe Institute of Technology, Karlsruhe, Germany  
stamatak@ics.forth.gr

## Abstract

Working with cognate data involves handling synonyms, that is, multiple words that describe the same concept in a language. In the early days of language phylogenetics it was recommended to select one synonym only. However, as we show here, binary character matrices, which are used as input for computational methods, do allow for representing the entire dataset including all synonyms. Here we address the question how one can and if one should include all synonyms or whether it is preferable to select synonyms a priori. To this end, we perform maximum likelihood tree inferences with the widely used RAxML-NG tool and show that it yields plausible trees when all synonyms are used as input. Furthermore, we show that a priori synonym selection can yield topologically substantially different trees and we therefore advise against doing so. To represent cognate data including all synonyms, we introduce two types of character matrices beyond the standard binary ones: probabilistic binary and probabilistic multi-valued character matrices. We further show that it is dataset-dependent for which character matrix type the inferred RAxML-NG tree is topologically closest to the gold standard. We also make available a Python interface for generating all of the above character matrix types for cognate data provided in CLDF format.

## 1 Introduction

Lexical data are frequently used as input to infer language trees via standard phylogenetic methods. Lexical data are typically cognate data that rely upon concept or meaning lists, such as the Swadesh List (Swadesh, 1955), for instance. When assembling data for a specific language, linguists typically attempt to identify a frequently used everyday word for each concept, which describes it as precisely as possible (Dunn, 2013). This data assembly process induces intrinsic difficulties, due to the inherent subjectivity in concept interpretation (List, 2018). Further, there often exist synonyms, that is, multiple words describe the same concept in a language and are used interchangeably by speakers. In German, for example, there are two words describing the concept "to kill": "töten" (related to the English word "dead") and "umbringen" (related to the English word "to bring"). The intricate nuances in meaning and usage are hard to determine and quantify (List, 2018). In the early days of language phylogenetics, Swadesh recommended to use the most common word only, "avoiding the complication of having to deal with a choice" (Swadesh, 1951). In "The ABC's of Lexicostatistics" (Gudschinsky, 1956), Gudschinsky advises to ensure objectivity by tossing a coin to decide which word to pick when there are several choices.

Computational phylogenetic methods have recently been applied to several cognate datasets. The inferences have mainly been conducted using Bayesian Inference (BI) methods (Kolipakam et al., 2018; Sagart et al., 2019; Heggarty, 2023), but a publication by (Jäger, 2018) shows that Maximum Likelihood (ML) based tree inference is useful as well, especially on extremely large language trees. For ML and BI approaches, that both heavily rely on the same type of phylogenetic likelihood calculations, cognate data are typically encoded via binary character matrices that represent the complete dataset including *all* synonyms. A question that has not been addressed to date is whether phylogenetic likelihood models as used in standard ML and BI tree inference can accommodate this data representation or whether it is preferable to choose synonyms via a labor-intensive, potentially error-prone, and subjective manual process beforehand.

Here, we focus on Maximum Likelihood (ML) tree inferences using the widely used RAxML-NG tool. Initially, working with empirical language data, we show that the topology of a tree inferred for a cognate dataset containing selected synonyms only can differ by up to 100% from the tree topology inferred for the corresponding complete dataset including all synonyms. Given these large potential discrepancies, we advise against manual selection. To alleviate this issue, we explore the potential benefits of using two types of alternative character matrices beyond the standard binary ones. The character matrices we propose can seamlessly be read as input by RAxML-NG while representing the complete dataset including all synonyms. We analyze the tree topologies resulting from ML inferences on all three character matrix types. We find that it depends on the respective dataset for which character matrix type the inferred tree best corresponds to the gold standard.

The remainder of this paper is structured as follows: First, we introduce our materials and methods. In particular, we formally define cognate data and describe the assembly process of the different character matrix types. Then, we evaluate how synonym selection affects the ML tree topologies inferred with RAxML-NG. Finally, we compare the introduced character matrix types. We consider the trees inferred with RAxML-NG and assess the matrix types based on how close the respective inferred trees are to the Glottolog gold standard.

## 2 Materials and Methods

### 2.1 Cognate Data

Each cognate dataset is based on a list of concepts. Collecting data for the languages under study results in an assignment of a set of words to each language-concept pair. From these data, we construct a matrix  $M$  containing the words' cognate classes. Cognate classes unite words that have been derived from a common ancestor (Dunn, 2013) (see Figure 1). We assume that the concept lists are reasonably assembled, that is, there exists at least one word for each concept in each language. When no word is given for a language-concept pair in the original data we interpret this as missing information.

If there are multiple synonym words describing a concept in a language, we say that the respective matrix cell is a *multi-state cell*. Otherwise, if there only is a single word describing a concept in a language, we call the respective matrix cell a *single-state cell*.

	big		big
English	big	E	big_1
English	great	E	big_2
German	groß	G	big_2
Dutch	groot	D	big_2
Norwegian	stor	N	big_3
Swedish	stor	S	big_3
	(a)		(b)

Figure 1: (a): Native cognate data (b): Corresponding matrix  $M$  with cognate classes (b)

### 2.2 Character Matrix Types

#### 2.2.1 Binary Character Matrices

A cognate dataset can be represented by a binary character matrix  $A^b$  containing the symbols 0 and 1). Additionally, specific entries may be set to the undetermined character  $-$ , to represent missing information. Given  $-$  at a certain column for a language, this language does not contribute anything to the respective per-column likelihood score in RAxML-NG. Hence, the missing entries do not affect the inference. However, the lack of information itself may impact the results (Roure et al., 2012).

We obtain  $A^b$  as the presence-absence-matrix corresponding to the matrix containing the cognate classes (see Figure 2 (a)). Each concept is therefore represented by as many columns as there are cognate classes, each corresponding to a specific cognate class. If there exists a word belonging to this cognate class in a certain language, the respective entry is set to 1, and to 0 otherwise. Thereby we assume that for each concept, there exists at least one word in every language. If there is no cognate class provided for a languages and a concept, this corresponds to missing information. We consequently set all columns corresponding to this concept to  $-$ .

### 2.2.2 Multi-valued Character Matrices

Some cognate datasets can also be represented by a multi-valued character matrix  $A^m$  containing multiple distinct symbols. In RAxML-NG, multi-valued character matrices are restricted to a maximum of 64 distinct symbols (Kozlov et al., 2019). In  $A^m$ , each concept is represented via a single data column only, but a different symbol is used for each cognate class. Multi-valued character matrices are thus restricted to represent a single cognate class for each language-concept pair. In order to construct  $A^m$ , the matrix  $M$  describing the respective cognate dataset must therefore contain single-state cells only. As this is generally not the case, we exclude multi-valued character matrices from further considerations.

	big_1	big_2	big_3		big
E	1	1	0	E	0, 1
G	0	1	0	G	1
D	0	1	0	D	1
N	0	0	1	N	2
S	0	0	1	S	2

(a)                      (b)

Figure 2: (a): Binary character matrix  $A^b$  (b): Multi-valued character matrix  $A^m$ , both corresponding to the cognate dataset in Figure 1. Note that  $A^m$  is invalid, because  $M$ (English, big) is a multi-state cell.

### 2.2.3 Probabilistic Character Matrices

The character matrices described so far are all deterministic, because we assume that a fixed symbol is observed at each data column for each language. In a probabilistic character matrix, we instead as-

sume that distinct symbols can occur with a certain probability, which is provided in the matrix. To represent missing data we explicitly set the probabilities for all symbols to 1.0 (Kozlov et al., 2019). This encoding does not provide any information, and hence, the missing entries do not contribute to the likelihood score.

We can represent a probabilistic character matrix in a file via the so-called CATG-Format that is supported by RAxML-NG (Kozlov et al., 2019). A tree inference based on a probabilistic character matrix differs from a standard inference with respect to the form of the conditional likelihood vectors at the tips. Usually, such a vector contains a single 1.0 entry for the observed discrete character while the remaining entries are all set to 0.0. In contrast, for a probabilistic character matrix, the conditional likelihood vectors are determined based on the provided probabilities (Kozlov et al., 2019) (see Figure 3).

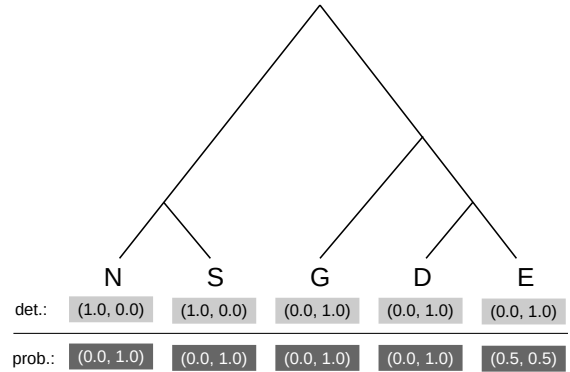


Figure 3: Tree with conditional likelihood tip vectors for the per-column likelihood of the column big\_2. The light gray vectors refer to the inference on  $A^b$  (Figure 2), the dark gray ones to the inference on  $A^{b*}$  (Figure 4).

When interpreting cognate data in a probabilistic manner, we can represent the datasets via probabilistic character matrices. If  $k$  synonyms exist for a concept in a language, we can assume that each of them occurs with probability  $\frac{1}{k}$ . Based on these probabilities, we can subsequently assemble probabilistic binary and multi-valued character matrices. In the probabilistic binary character matrix  $A^{b*}$ , a concept is represented by as many columns as there are cognate classes, just as in the corresponding deterministic binary character matrix  $A^b$ . At a column corresponding to one of the  $k$  synonyms, we observe the symbol 1 with probability  $\frac{1}{k}$  and the symbol 0 with probability  $1 - \frac{1}{k}$  for the respective

language. In the multi-valued probabilistic character matrix  $A^{m*}$ , each cognate class related to a certain concept is encoded by a different symbol but the concept is represented by only one single column. At this column, we observe each of the symbols representing one of the  $k$  synonyms with probability  $\frac{1}{k}$  (see Figure 4).

	big_1	big_2	big_3
E	(0.5, 0.5)	(0.5, 0.5)	(1.0, 0.0)
G	(1.0, 0.0)	(0.0, 1.0)	(1.0, 0.0)
D	(1.0, 0.0)	(0.0, 1.0)	(1.0, 0.0)
N	(1.0, 0.0)	(1.0, 0.0)	(0.0, 1.0)
S	(1.0, 0.0)	(1.0, 0.0)	(0.0, 1.0)

(a)

	big
E	(0.5, 0.5, 0.0)
G	(0.0, 1.0, 0.0)
D	(0.0, 1.0, 0.0)
N	(0.0, 0.0, 1.0)
S	(0.0, 0.0, 1.0)

(b)

Figure 4: (a): Probabilistic binary character matrix  $A^b$  (b): Probabilistic multi-valued character matrix  $A^m$ , both corresponding to the cognate dataset in Figure 1.

### 2.3 Comparing Trees

We measure topological dissimilarities between inferred phylogenetic trees using the *Robinson-Foulds (RF) distance* (Robinson and Foulds, 1981). This standard metric is based on non-trivial splits in trees. A split is a partitioning of the tree’s tips into two sets corresponding to the subtrees that arise when a branch of the tree is removed. A split is called non-trivial, if each set contains at least two tips. The absolute RF distance between two trees is defined as the number of non-trivial splits, which are unique to one of the two trees. In the following, we use the relative RF distance which we obtain by normalizing the absolute RF distance with  $2(n-3)$ , the total number of non-trivial splits in two strictly binary unrooted trees.

The inferred ML trees are strictly binary, but polytomies can occur in manually constructed reference trees (e.g., in Glottolog reference trees, see below). To compare an inferred ML tree to a reference tree, we therefore use the *generalized quartet (GQ) distance* (Pompei et al., 2011) instead. This metric

has the advantage that it yields a distance of 0 if there are no contradictions between the inferred tree and the reference tree, even if the reference tree does contain polytomies. To calculate the GQ distance, we extract all possible quartets of tips in the tree. For each quartet, we then determine the topology of the 4-tip subtree induced by the comprehensive tree. When comparing two trees, the GQ reflects the proportion of quartets that induce distinct topologies. Note that the RF distance and the GQ distance are distributed differently (Steel and Penny, 1993), which is illustrated by the following example: Let  $T$  be a fully balanced strictly binary tree with 16 leaves and let further  $T'$  be a tree obtained from  $T$  by swapping two leaves from neighboring subtrees (for details refer to Figure 11 in the supplement). The RF distance of  $T$  and  $T'$  is 0.15, while they exhibit a GQ distance of 0.03 only.

### 2.4 Maximum Likelihood Tree Inferences

In our experiments, we execute 20 independent ML tree searches on each character matrix under study. We use the default tree search configuration of RAxML-NG (10 searches starting from random trees and 10 searches starting from randomized stepwise addition order parsimony trees). For the inferences on both, the deterministic, and the probabilistic binary character matrices, we use the BIN+G model of binary character substitution. For the tree searches on probabilistic multi-valued character matrices we use the MK+G model which allows using up to 64 different characters but assumes equal substitution rates between all characters. Using BIN+G (MK+G resp.), we approximate the  $\Gamma$  model of rate heterogeneity via four discrete rates. Thus, each inference includes the ML estimation of the  $\alpha \in [0, 100]$  shape parameter that determines the shape of the  $\Gamma$  distribution. The smaller the estimate of  $\alpha$ , the higher the rate heterogeneity in the respective dataset will be (Yang, 1995). The command lines we used to execute the inferences are available on Github (<https://github.com/luisevonderwiese/synonyms>).

### 2.5 Quantifying Difficulty

To quantify the difficulty of a phylogenetic inference for a given dataset we use the difficulty score as predicted by Pythia (Haag et al., 2022). The tool internally uses a Random Forest Regressor (Ho, 1995) to predict this difficulty score based on attributes of the character matrices and on the

results of computationally substantially less expensive parsimony-based tree inferences (Farris, 1970). Because the parsimony approach can only be applied to deterministic character matrices, the difficulty score is also limited to this matrix type.

## 2.6 Data

For our analyses we use 44 cognate datasets. We retrieve the vast majority (39 datasets) from the cross-linguistic lexical database *Lexibank* (List et al., 2022). The five remaining datasets originate from the supplementary material provided for the book "Sequence Comparison in Historical Linguistics" (List, 2021). The above repositories comprise more datasets than we use here, as not all of them are suitable for our experiments. Since GQ distances can only be calculated on trees with strictly more than 4 tips, we exclude datasets comprising less than 5 languages. We also do not consider datasets with more than 400 languages due to the excessive tree inference times. In addition, these datasets exhibit an unfavorable number of concepts to number of languages ratio which yield them difficult to reliably infer. We therefore do not expect the respective tree inference results to be informative. We further exclude datasets with regard to the maximum number of different cognate classes that are occurring for the concepts. If only one cognate class is used for each concept, we do not consider the respective dataset, as the corresponding character matrices are not informative. We also exclude datasets comprising concepts with more than 64 distinct cognate classes, as no probabilistic multi-valued character matrix can be constructed in this case, because RAxML-NG is limited to using a maximum of 64 distinct symbols.

As gold standard, we use the manually constructed tree published by *Hammarström et al.* in the *Glottolog* database. This tree contains all 8205 languages listed in Glottolog. For each dataset, we obtain the respective gold standard tree by constraining the comprehensive tree to the languages contained therein. For a dataset to be suitable for our experiments, it must be possible to extract an informative reference tree. To this end, we exclude datasets, if the corresponding reference tree has a star topology because comparisons of binary topologies with a star topology do not yield meaningful topological distances.

In both data sources, the datasets are standardized as specified by the Cross-Linguistic Data

Format (CLDF) (Forkel et al., 2018). Our implementation for converting CLDF data into the character matrix types we describe here is available on Github (<https://github.com/luisevonderwiese/lingdata>).

## 3 Results

### 3.1 Effects of Synonym Selection

In the following, we aim to assess, whether ML tree inferences are feasible on binary character matrices representing cognate datasets with *all* synonyms or whether it is preferable to choose synonyms manually in advance. For this purpose, we investigate how it impacts the results of RAxML-NG based tree inferences when different combinations of synonyms are being selected. The experimental setup is illustrated in Figure 5

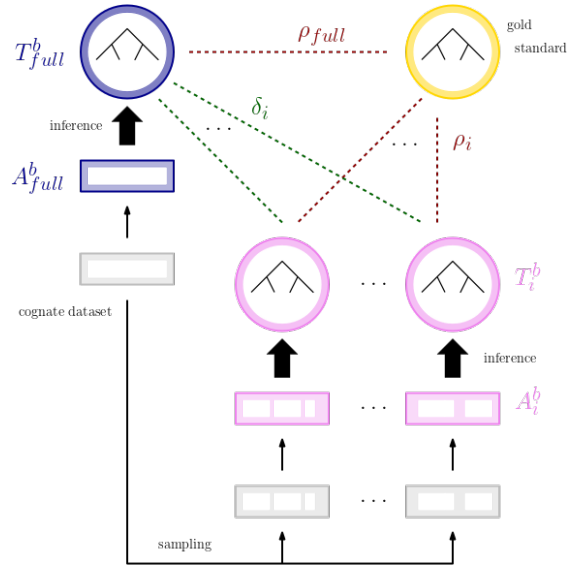


Figure 5: Experimental setup for assessing the effects of synonym selection: For each dataset, we create 100 selection samples and construct the corresponding (deterministic) binary character matrices  $A_i^b$ ,  $i \in 1, \dots, 100$ .  $A_{full}^b$  represents the complete dataset including all synonyms. For each character matrix we consider the best scoring tree resulting from 20 independent tree searches with RAxML-NG, denoted by  $T_i$ ,  $i \in 1, \dots, 100$ ,  $T_{full}$  respectively.  $\delta_i$  corresponds to the RF distance between  $T_i$  and  $T_{full}$ ,  $\rho_i$  to the GQ distance between  $T_i$  and the gold standard tree. By  $\rho_{full}$  we denote the GQ distance between  $T_{full}$  and the gold standard.

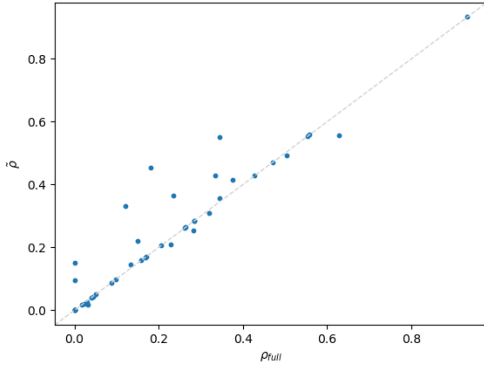


Figure 6: Each marker corresponds to a specific dataset under study. The x-axis indicates  $\rho_{\text{full}}$ , the GQ distance of the respective tree  $T_{\text{full}}$  to the gold standard. The y-axis indicates  $\tilde{\rho}$ , the median of the GQ distances of the trees  $T_i$  to the gold standard. For the majority of the datasets, the tree  $T_{\text{full}}$  is at least as close to the gold standard as the median of the trees  $T_i$ .

We generate selection samples from each examined dataset by selecting synonyms uniformly at random. Hence, if  $k$  synonyms exist for a concept in a language, we pick one of them with probability  $\frac{1}{k}$ . We create 100 such selection samples for each dataset under study. For some of the datasets, we performed the following experiments with 1000 samples, but this did not lead to substantial differences in the results. For each sample we construct a (deterministic) binary character matrix  $A_i^b$ ,  $i \in 1, \dots, 100$ . Note that each of these matrices only contains information about the selected synonyms. Then, using RAxML-NG with the tree search options as outlined above we execute 20 independent ML tree searches on each of these character matrices as well as on the character matrix  $A_{\text{full}}^b$  representing the complete dataset including all synonyms. For each character matrix  $A_i^b$ , we consider the best-scoring ML tree  $T_i$  we inferred on it. Let further  $T_{\text{full}}$  be the best-scoring tree resulting for  $A_{\text{full}}^b$ . For each dataset, we use the corresponding gold standard tree from Glottolog as a reference. Let  $\rho_{\text{full}}$  be its GQ distance to  $T_{\text{full}}$ . For each tree  $T_i$ , we denote its GQ distance to the gold standard by  $\rho_i$ . Let further  $\tilde{\rho}$  be the median of the GQ distances  $\rho_i$ . For  $\rho_{\text{full}}$ , we obtain an average distance of 0.22 over all 44 datasets while  $\tilde{\rho}$  averages to 0.25. Thus, the two approaches appear to perform equally well at first sight, with  $T_{\text{full}}$  being only slightly closer to the reference, on average.

For a more detailed assessment, we compare  $\rho_{\text{full}}$  and  $\tilde{\rho}$  for each dataset (see Figure 6). For 33 out of 44 examined datasets, we observe that  $\rho_{\text{full}} \leq \tilde{\rho}$ , that is  $T_{\text{full}}$  comes closer to the gold standard. In most of the cases, the inference on  $A_{\text{full}}^b$  thus performs better than the median inference on the sampled character matrices. For the datasets where  $\rho_{\text{full}} > \tilde{\rho}$  applies, the difference never exceeds 0.07 and only for 5 datasets it exceeds 0.01. If the median tree  $T_i$  is closer to the gold standard, the differences are hence not substantial in most cases. These results speak in favor of performing inferences on the full dataset as the results tend to be slightly better than for the median randomized synonym selection. When using the mean GQ distance instead of the median, the observations are analogous. For details refer to Figure 12 in the supplement.

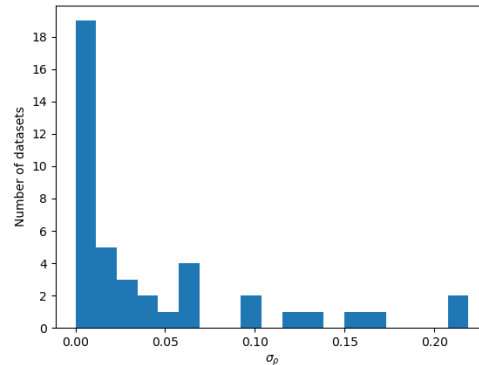


Figure 7: Distribution of  $\sigma_\rho$ , the standard deviation of the GQ distances of the trees  $T_i$  to the gold standard. The x-axis indicates  $\sigma_\rho$ , the y-axis the number of datasets such that respective standard deviation occurs. The observed standard deviations indicate that the inferred trees are substantially different depending on the subset of selected synonyms.

The advantages of using the full datasets are further emphasized when analyzing the stability of the tree inference under the described sampling procedure. For this purpose, we evaluate how much the trees  $T_i$  vary in terms of their deviation from the gold standard. For a fixed dataset, let  $\sigma_\rho$  be the standard deviation of the GQ distances  $\rho_i$ . The distribution of  $\sigma_\rho$  over all examined datasets is depicted in Figure 7. For 13 datasets, we observe  $\sigma_\rho > 0.05$ , for 2 of them, even  $\sigma_\rho > 0.2$ . Since the GQ distance is

distributed differently than the RF distance, the observed standard deviations indicate substantial differences. How close the inferred tree comes to the gold standard can therefore vary considerably depending on the subset of selected synonyms.

In an additional stability analysis, we examine, to which extent the trees  $T_i$  deviate from the respective tree  $T_{\text{full}}$ . For each tree  $T_i$ , we therefore determine  $\delta_i$  as its RF distance to  $T_{\text{full}}$ . We consider  $\bar{\delta} := (\sum_{i=1}^{100} \delta_i)/100$ , indicated on the x-axis of Figure 8. The figure's y-axis gives the number of examined datasets for which the respective average distance results from the described experiment.  $\bar{\delta}$  is close to 0 for some datasets but also  $\geq 0.4$  or even  $\geq 0.8$  for a considerable proportion of datasets.  $\bar{\delta}$  can be interpreted as an indication for the stability of the inferred tree under sampling. The lower  $\bar{\delta}$ , the more stable the respective dataset is.

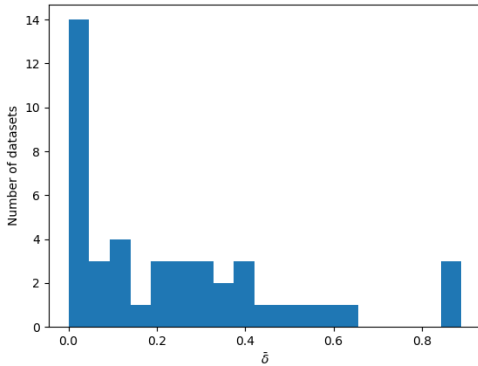


Figure 8: Distribution of  $\bar{\delta} := (\sum_{i=1}^{100} \delta_i)/100$ . The x-axis indicates  $\bar{\delta}$ , the y-axis the number of datasets such that the trees  $T_i$  yield the respective average RF distance to  $T_{\text{full}}$ . The lower  $\bar{\delta}$ , the more stable the respective dataset is under random synonym selection.

For each dataset we also determine the difficulty of the inference on the character matrix  $A_{\text{full}}^b$ . The obtained difficulty score is only slightly correlated with the stability as quantified by  $\bar{\delta}$  (Pearson correlation 0.43, P-value 0.003). Further, we examine the proportion of multi-state cells for each cognate dataset, corresponding to the proportion of language-concept pairs such that there are words from more than one cognate class describing the concept in the language. This score is slightly correlated with  $\bar{\delta}$  (Pearson correlation

0.45, P-value 0.002). The more multi-state cells exist, the more information are discarded during sampling and the more the sampled character matrix differs from the character matrix that is based on the full dataset, leading to the observed correlation.

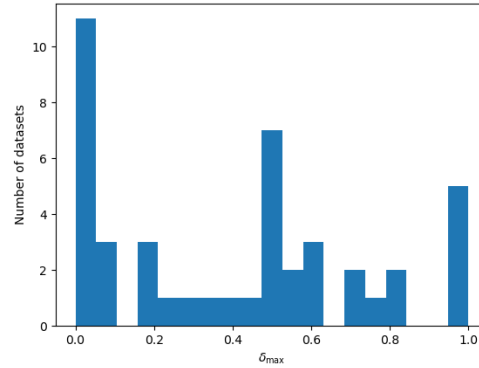


Figure 9: Distribution of  $\delta_{\max} := \max(\delta_i : i \in 1, \dots, 100)$ .

The x-axis indicates  $\bar{\delta}$ , the y-axis the number of datasets such that respective maximum value occurs among the RF distances of the trees  $T_i$  to  $T_{\text{full}}$ . The analysis illustrates that there exist datasets for which the worst case synonym choice leads to a tree which is entirely different (RF distance of 1.0) from the tree inferred on the dataset including all synonyms.

We finally consider  $\delta_{\max} := \max(\delta_i : i \in 1, \dots, 100)$ . The analysis of this score elucidates the detrimental effect of an unfavorable worst case synonym choice. Figure 9 depicts the respective distribution for the datasets under study. We observe that there is a considerable proportion of datasets with  $\delta_{\max} \geq 0.5$ . For 5 datasets, we even observe  $\delta_{\max} = 1$ . For these datasets there exists at least one character matrix containing a certain subset of synonyms such that the inferred tree admits an RF distance of 1.0 to the tree resulting from the full dataset.

Our observations illustrate that the synonym selection can induce entirely different RAXML-NG based ML tree topologies. The decision, which synonyms to consider thus substantially affects the results. Therefore, we strongly advise against manual synonym selection. Instead, we recommend to consider all known synonyms when inferring phylogenetic trees. Our analysis shows, that tree inference on the respective representation as a binary

character matrix leads to feasible results. Additionally, this circumvents the labor-intensive process of selecting synonyms manually.

### 3.2 Modeling Data with Synonyms

In the following, we compare the performance of ML tree inferences on three different kinds of character matrices representing cognate data. For each dataset under study, we consider its representation as a deterministic binary character matrix  $A^b$ , as a probabilistic binary character matrix  $A^{b*}$ , and as a probabilistic multi-valued character matrix  $A^{m*}$ . On each character matrix type we execute 20 independent ML tree searches with RAxML-NG as described above. The structure of the following experiment is illustrated in Figure 10.

We aim to assess how suitable the different character matrix types are for capturing the signal contained in the data during ML tree inference with RAxML-NG. To this end, we examine the resulting trees. For a fixed dataset, let  $T^b$ ,  $T^{b*}$ ,  $T^{m*}$  be the best-scoring tree inferred on the respective character matrix type. We compare these trees to the corresponding gold standard from Glottolog. We are interested in which character matrix type will induce the tree that is closest to it. We henceforth say that this is the character matrix type performing best.

The trees inferred on  $A^b$  yield an average GQ distance of 0.22 to the respective gold standard tree. For both probabilistic character matrix types we observe an average GQ distance of 0.23. At first glance, all character matrix types therefore appear to yield results of comparable quality. For 9 datasets,  $A^b$  performs best, for 9 datasets it is  $A^{b*}$ , and for 11 datasets  $A^{m*}$ . In 8 datasets, all character matrix types perform equally well. There are 7 datasets where two distinct character matrix types yield equally good results, but are better than the third. For the sake of simplicity, we exclude these datasets from our further analyses. There is no clear trend for one character matrix type always being preferable over the others. Conversely, based on our results, we cannot advise against the use of any type of character matrix.

Subsequently, we analyze the differences between the trees inferred on the different character matrix types, aiming to show that these differences are so substantial that it is indeed worthwhile to

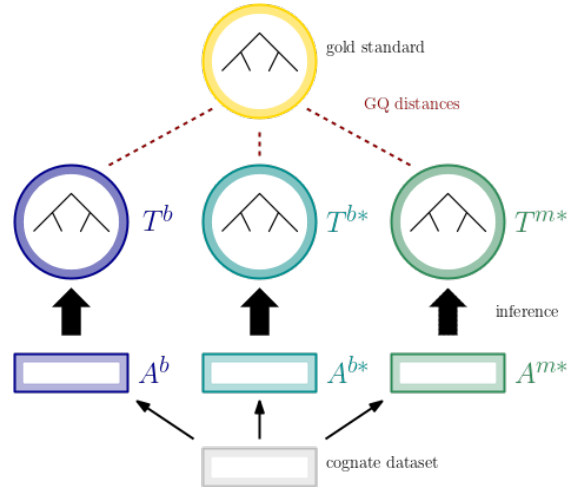


Figure 10: Experimental setup for comparing the performance of different character matrix types: For each dataset we construct a deterministic binary character matrix  $A^b$ , a probabilistic binary character matrix  $A^{b*}$ , and a probabilistic multi-valued character matrix  $A^{m*}$ . For each character matrix type, we consider the best tree scoring tree resulting from 20 independent tree searches with RAxML-NG, denoted by  $T^b$ ,  $T^{b*}$ ,  $T^{m*}$  respectively. We compute the GQ distances of these trees to the gold standard from Glottolog.

perform inferences on all three character matrix types. For this purpose we first consider the datasets, for which  $A^b$  performs best. Over these datasets,  $T^b$ , which comes closest to the gold standard tree, yields an average RF distance of 0.60 to  $T_d^{b*}$  and of 0.61 to  $T_d^{m*}$ . This indicates that the trees resulting from the inferences on the various character matrix types can differ considerably. We observe an analogous behavior for the datasets where  $A^{b*}$  or  $A^{m*}$  perform best (see Table 1 in the supplement). We conclude that the differences between the inferred trees can become so large such that all three character matrix types always need to be considered.

In the following, we attempt to identify properties, that for a given dataset, might be able to predict, which character matrix type performs best. First, we consider the  $\alpha$  shape parameter that determines the shape of the  $\Gamma$  distribution under the BIN+G (MK+G resp.) model. For 9 out of the 11 datasets with  $A^{m*}$  performing best, the ML estimates of  $\alpha$  are below 20, indicating a moderate to high degree of rate heterogeneity. This is also the case for 8 out of the 9 datasets for which  $A^{b*}$  performs best. However,  $\alpha < 20$  is only observed



for 5 out of the 9 datasets for which the binary character matrix yields the best performance and only for 4 out of the 8 datasets for which all character matrix types perform equally well. We therefore observe a tendency for probabilistic modeling to be advantageous for datasets with high rate heterogeneity. These datasets may exhibit a larger variance with respect to the number of cognate classes per concept, which can possibly be better accommodated via a probabilistic character matrix type.

We further investigate whether the difficulty of the ML inference on a certain dataset is related to character matrix performance. In the following, the difficulty reported for a dataset is the difficulty obtained for  $A^b$ , as difficulty scores can only be computed for deterministic character matrices. The datasets for which all character matrix types perform equally well exhibit a comparably low average difficulty of 0.17. A low difficulty score indicates a strong phylogenetic signal in the data. This strong signal can be captured during the ML tree inference, regardless of the type of character matrix used to represent it. The datasets with  $A^b$  performing best exhibit an average difficulty of 0.45. For the datasets for which  $A^{b*}$  performs best, the average difficulty is 0.29, and for those for which  $A^{m*}$  yields the best performance, it amounts to 0.18. While the probabilistic character matrix types are hence advantageous for data with a clear signal, the deterministic representation is better suited to capture the signal in datasets where this is more difficult.

## 4 Conclusion

We demonstrated, that the selection of synonyms in cognate datasets *can* induce substantially different tree topologies when performing ML inferences with RAxML-NG. It is thus preferable to perform tree inferences on the full dataset with all synonyms included. This also circumvents the labor-intensive process of manual synonym selection. The datasets can be encoded as (deterministic) binary character matrices. In addition, we introduced probabilistic binary and probabilistic multi-valued character matrices as alternative representations. We showed that it is dataset-dependent, for which character matrix type the inferred tree is closest to the gold standard. We were able to identify the rate heterogeneity and the difficulty score as properties

that may indicate which character matrix type is best suited for a given dataset. Note that unfortunately, the number of available cognate datasets is too low in order to train any machine-learning based predictors. We therefore recommend performing inferences on all three character matrix types when analyzing cognate datasets. We provide a Python interface on Github (<https://github.com/luisevonderwiese/lingdata>) that can be used to create all of the above character matrices for any cognate dataset provided in CLDF format.

## 5 Future Work

Our work leads to several novel questions that need to be addressed. When constructing probabilistic character matrices, data from corpus studies could be taken into account instead of assuming a uniform probability distribution of synonym occurrence, albeit this information could be challenging to obtain for small languages and dialects. Future work should further strive to develop an alternative model of evolution taking idiosyncrasies of language data into account. Another open question is how the quality of the inference methods can be assessed without referring to the gold standard. This requires the development of language-specific data simulation tools, taking into account the challenges that have been described with respect to simulating realistic DNA data (Troost et al., 2023).

## Acknowledgement

Luise Häuser and Alexandros Stamatakis are financially supported by the Klaus Tschira Foundation, and by the European Union (EU) under Grant Agreement No 101087081 (CompBiodiv-GR). Gerhard Jäger is supported by the ERC-AdG 834050 (CrossLingference) and the DFG-FOR 2234 *Words, Bones, Genes, Tools*.

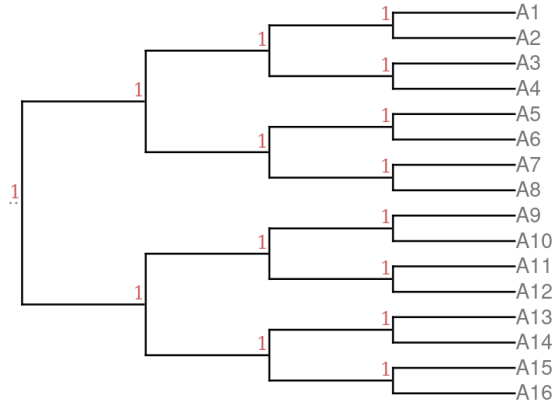


Funded by  
the European Union

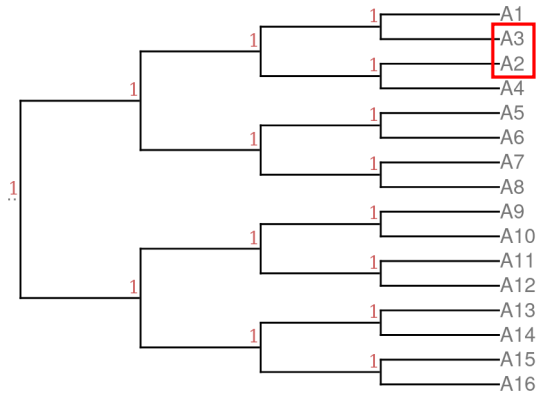
## References

- Michael Dunn. 2013. Language phylogenies.
- James S. Farris. 1970. Methods for computing wagner trees. *Systematic Zoology*, 19(1):83–92.
- Robert Forkel, Johann-Mattis List, Simon Greenhill, Christoph Rzymiski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon Kaiping, and Russell Gray. 2018. [Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics](#). *Scientific Data*, 5:180205.
- Sarah C. Gudschinsky. 1956. The abc’s of lexicostatistics (glottochronology). *WORD*, 12:175–210.
- Julia Haag, Dimitri Höhler, Ben Bettisworth, and Alexandros Stamatakis. 2022. [From easy to hopeless - predicting the difficulty of phylogenetic analyses](#). *bioRxiv*.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2022. [Glottolog 4.7](#).
- Paul et al. Heggarty. 2023. [Language trees with sampled ancestors support a hybrid model for the origin of indo-european languages](#). *Science*, 381(6656).
- Tin Kam Ho. 1995. [Random decision forests](#). In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1.
- Gerhard Jäger. 2018. [Global-scale phylogenetic linguistic inference from lexical resources](#). *Scientific Data*, 5.
- Vishnupriya Kolipakam, Fiona M. Jordan, Michael Dunn, Simon J. Greenhill, Remco Bouckaert, Russell D. Gray, and Annemarie Verkerk. 2018. [A Bayesian phylogenetic study of the Dravidian language family](#). *Royal Society Open Science*, 5(171504):1–17.
- Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. 2019. [RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference](#). *Bioinformatics*, 35(21):4453–4455.
- Johann-Mattis List, Robert Forkel, Simon Greenhill, Christoph Rzymiski, Johannes Englisch, and Russell Gray. 2022. [Lexibank, a public repository of standardized wordlists with computed phonological and lexical features](#). *Scientific Data*, 9:316.
- Mattis List. 2018. [Tossing coins: linguistic phylogenies and extensive synonymy](#).
- Mattis List. 2021. *Sequence Comparison in Historical Linguistics*. düsseldorf university press, Berlin, Boston.
- Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. [On the accuracy of language trees](#). *PloS one*, 6:e20109.
- D.F. Robinson and L.R. Foulds. 1981. [Comparison of phylogenetic trees](#). *Mathematical Biosciences*, 53(1):131–147.
- Béatrice Roure, Denis Baurain, and Hervé Philippe. 2012. [Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets](#). *Molecular Biology and Evolution*, 30(1):197–214.
- Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. [Dated language phylogenies shed light on the ancestry of sino-tibetan](#). *Proceedings of the National Academy of Science of the United States of America*, 116:10317–10322.
- Mike Steel and David Penny. 1993. [Distributions of tree comparison metrics—some new results](#). *Systematic Biology - SYST BIOL*, 42:126–141.
- Morris Swadesh. 1951. [Diffusional cumulation and archaic residue as historical explanations](#). *Southwestern Journal of Anthropology*, 7(1):1–21.
- Morris Swadesh. 1955. [Towards greater accuracy in lexicostatistic dating](#). *International Journal of American Linguistics*, 21(2):121–137.
- Johanna Trost, Julia Haag, Dimitri Höhler, Laurent Jacob, Alexandros Stamatakis, and Bastien Boussau. 2023. [Simulations of Sequence Evolution: How \(Un\)realistic They Are and Why](#). *Molecular Biology and Evolution*, 41(1):msad277.
- Z Yang. 1995. [A space-time process model for the evolution of dna sequences](#). *Genetics*, 139(2):993–1005.

## A Supplementary Figures and Tables



(a) Tree  $T$ , fully balanced strictly binary tree with 16 leaves



(b) Tree  $T'$ , differences to  $T$  are highlighted in red

Figure 11: The trees  $T$  and  $T'$  exhibit an RF distance of 0.15 but a GQ distance of 0.03, which illustrates the different distributions of the metrics.

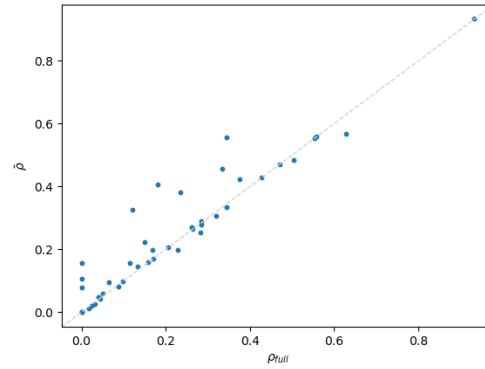


Figure 12: Each marker corresponds to a specific dataset under study. The x-axis indicates  $\rho_{full}$ , the GQ distance of the respective tree  $T_{full}$  to the gold standard. The y-axis indicates  $\bar{\rho}$ , the average GQ distance of the trees  $T_i$  to the gold standard. Most markers are located on the identity (represented by a dashed line) or above, that is the tree  $T_{full}$  is at least as close to the gold standard as the average tree  $T_i$  for the majority of the datasets.

Mean RF( $\cdot, \cdot$ )	$T^b$	$T^{b*}$	$T^{m*}$
Datasets with $A^b$ performing best:			
$T^b$	0.00	0.60	0.61
Datasets with $A^{b*}$ performing best:			
$T^{b*}$	0.25	0.00	0.36
Datasets with $A^{m*}$ performing best:			
$T^{m*}$	0.43	0.40	0.00

Table 1: Mean RF distances among the trees resulting from the inferences on the different character matrix types. Datasets are grouped according to the character matrix type performing best. The differences between the trees are so substantial that it is indeed worthwhile to perform inferences on all three character matrix types.