Karlsruhe Institute of Technology

Quantitative Analysis and Characterization of Natural Language Evolution Datasets

Master Thesis of

Luise Häuser

At the Department of Informatics
Institute of Theoretical Informatics

Reviewers:	Prof. Dr. Alexandros Stamatakis
	Prof. Dr. Thomas Bläsius
Advisor:	M.Sc. Julia Haag

16th January 2023 – 17th July 2023

Statement of Authorship

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, June 14, 2023

Abstract

Methods for phylogenetic inference have been developed mainly for the reconstruction of evolutionary relationships of species based on biological sequence data. However, these methods are also made use of in linguistics for inferring phylogenies concerning the evolution of natural languages. In the scope of this thesis, we examine the corresponding linguistic input data. We conduct a case study on an exemplary morphosyntactic data set, examining various methods to analyze the signal it contains and to eliminate geographical information the data may include. Further, we perform analyses on numerous linguistic data sets collected from various sources and assembled in a database. We compare these data sets to morphological data from biology, considering differences in the behavior of phylogenetic inferences with RAxML-NG. Additionally, we investigate how it impacts the tree inferences, whether we represent a data set by a binary or by a multi-valued MSA. We study how to model subjectivity related with synonym selection in cognate data. We present probabilistic MSAs as a possible solution and show on an example data set that this might be an appropriate approach.

Deutsche Zusammenfassung

Methoden zur phylogenetischen Inferenz wurden hauptsächlich entwickelt, um mithilfe biologischer Sequenzdaten die evolutionären Beziehungen zwischen Spezies zu rekonstruieren. Diese Methoden werden aber auch in der Linguistik eingesetzt um Stammbäume für die Evolution natürlicher Sprachen zu erhalten. Im Rahmen dieser Arbeit untersuchen wir die entsprechenden linguistischen Eingabedaten. Wir führen eine Fallstudie zu einem exemplarischen morphosyntaktischen Datensatz durch und untersuchen verschiedene Methoden, um das darin enthaltene Signal zu analysieren und eventuell enthaltene geographische Informationen zu eliminieren. Darüber hinaus analysieren wir zahlreiche linguistische Datensätzen, die aus verschiedenen Quellen stammen und aus denen wir eine Datenbank zusammengestellt haben. Wir vergleichen diese Datensätze mit morphologischen Daten aus der Biologie und betrachten die Unterschiede im Verhalten der phylogenetischen Inferenz mit RAxML-NG. Außerdem untersuchen wir, wie es sich auf die Inferenz auswirkt, ob wir einen Datensatz durch ein binäres oder durch ein multi-value MSA darstellen. Wir betrachten schließlich, wie die Subjektivität bei der Synonymauswahl in Kognat-Daten modelliert werden kann. Wir stellen probabilistische MSAs als mögliche Lösung vor und zeigen an einem Beispieldatensatz, dass dieser Ansatz geeignet erscheint.

Contents

List of Figures

1. Introduction

1.1. Motivation

Our planet is populated by a multitude of creatures. How this diversity, and with it mankind itself, came into existence is a fundamental question occupying us for centuries. Charles Darwin was the first to describe the basic mechanisms of evolution in his famous book "On the Origin of Species" published in 1859 [17]. He introduced the idea of a "tree of life", according to which all species evolved from a common ancestor by splitting and continuous modification (see Figure 1.1). Since then, scientists aim to reconstruct such evolutionary trees. In the recent decades, advances in molecular biology (e.g. next generation sequencing [5]) and computational methods [69] have spurred the development of new approaches to this end.
The languages spoken by humans are, as well, impressively diverse. According to data from *Hammarström et al.* [40] collected in 2022, there are 7636 different languages spoken by people as their first language. Again, this raises the question, how such a impressive variety emerged. In his book "The Descent of Man" (1871) [18] Darwin notes, that species and languages develop in a "curiously parallel" way. Concerning the evolution of languages, August Schleicher and Friedrich Schlegel are considered as the pioneers, using language evolution trees even before Darwin published "On the Origin of Species" [4]. A language evolution tree published later by Schleicher is depicted in Figure 1.2.

Species and languages can be considered as two concrete examples of *evolvable systems*. According to *Ladoukakis et al.* [45], a system is called evolvable, if it fulfills the following conditions: (1) it comprises populations of units, which are able to replicate, (2) replication goes along with the inheritance of characteristics, (3) characteristics vary, because during inheritance, there is a non-zero probability for mistakes. It is hence straight-forward to apply the methods for the reconstruction of evolutionary relationships of species to other evolvable systems as well. In the scope of this thesis, we pursue this approach with respect to the evolution of languages and specifically consider the data that are used as input for the computational methods. In the introductory part, we present all concepts relevant for our work. We keep the descriptions as general as possible and differentiate only where necessary by application in biology or linguistics. We briefly show where the differences between the evolution of species and languages lie and indicate possible related challenges for the inference methods used.

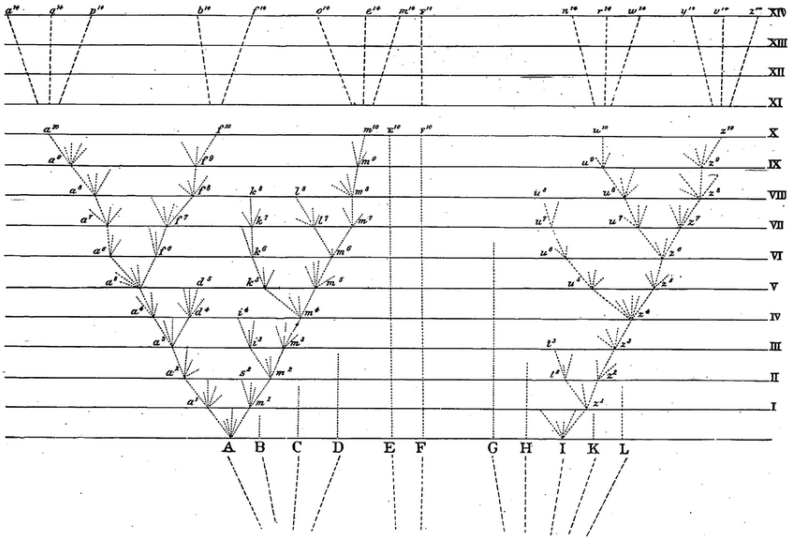


Figure 1.1.: Charles Darwin's "tree of life" (1859) [17]

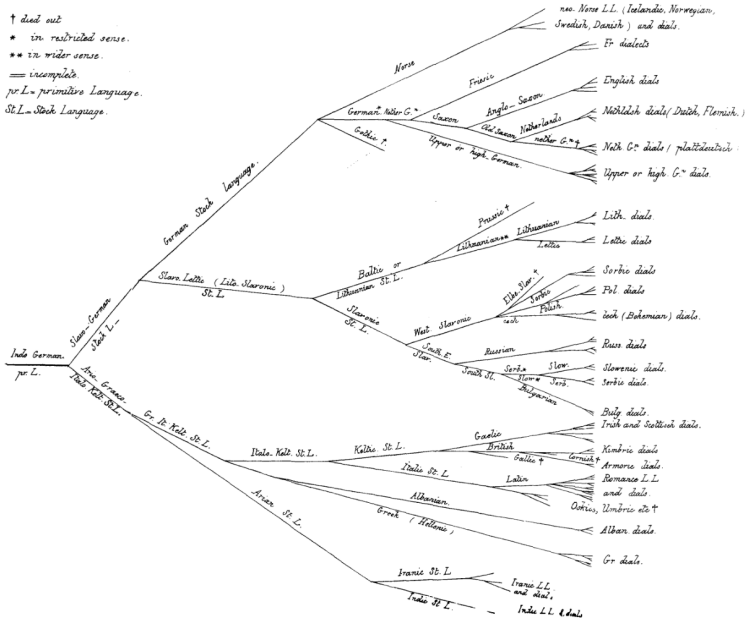


Figure 1.2.: August Schleicher's language evolution tree of the Indo-European language family (1863) [66]

1.2. Phylogenetic Trees

Let $\mathcal{T}$ be the set of units, whose evolution we intend to study. We call these units *taxa* (singular *taxon*). By $n \coloneqq |\mathcal{T}|$ we denote the number of examined taxa. A *phylogenetic tree* is a tree representing the hypothetical evolutionary relationships of a set of taxa $\mathcal{T}$. Each taxon is assigned to a leaf of the tree. *External* branches connect leaves with the tree, all remaining branches are *internal* branches. The branch lengths typically correspond to relative evolutionary distances. Phylogenetic trees are usually strictly bifurcating, that is all inner nodes have two child nodes. Further, phylogenetic trees are usually unrooted. However there do exist dedicated methods for rooting them (see also Section 2.2.3).

To measure topological dissimilarities between phylogenetic trees, we use the *Robinson-Foulds distance* (RF distance) [63]. This metric is based on splits in trees. A split is a partitioning of the taxa into two sets corresponding to the subtrees that arise when a branch of the tree is removed. A split is called non-trivial, if the respective branch is an internal one. The trivial splits are not of interest, since they all occur in every possible topology. The absolute RF distance of two trees is the number of non-trivial splits, which are induced by either one of the two trees but not by the other one. A binary tree with n taxa has $n-3$ internal branches, and hence, the maximum RF distance is $2(n-3)$. The relative RF distance of two trees is therefore defined as their absolute RF distance divided by $2(n-3)$. Unless otherwise stated, the given RF distances are relative. We further note that the RF distance is a metric which only captures dissimilarities in the trees' topologies and ignores differences in the branch lengths.

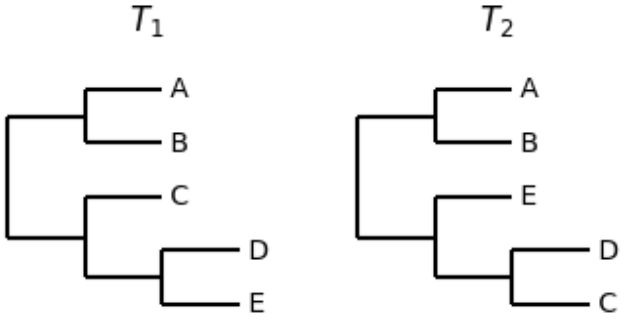


Figure 1.3.: RF Distance of T_1 and T_2 with $n = 5$
Splits in T_1: `AB|CDE`, `ABC|DE`, Splits in T_2: `AB|CDE`, `ABE|DC`
RF Distance: $\frac{2}{2(n-3)} = \frac{1}{2}$

1.3. Input data

Before the availability of large-scale DNA sequencing, lineage trees in biology were inferred based on *morphological data* [46] containing information about a specie's outer appearance as well as features describing the structure and relationships of internal parts [76]. Due to technical advance, it has been possible for several decades by now to obtain *sequence data*, representing the genetic material which encodes the observable traits [46].

Languages however lack this type of a logical code which is expressed in observable features. Analogous to morphological studies in biology, linguists examine structural properties of languages, resulting in *morphological linguistic data*. Additionally, they obtain data by investigating the vocabulary (*cognate data*) and the sounds (*sound class data*) occurring in languages.

We begin with the introduction of sequence DNA data and use it to explain the input format for phylogenetic inference. Morphological data from biology as well as the types of data used in linguistics can be summarized under the term *categorical data*. In the following, we describe sequential and categorical data for phylogenetics in more detail. We further explain how we transform categorical data in order to use it as an input for a phylogenetic inference.

1.3.1. Sequence Data

Sequences of Deoxyribonucleic Acid (DNA) encode the hereditary information of the species. DNA is a large molecule consisting of long strands of base pairs [3]. The sequence strings (reads) obtained from DNA sequencing contain symbols from $\Sigma = (\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T})$, each of them representing one of the four bases occurring in DNA [68]. Phylogenetic inference is based on homologous sequences observed in different organisms or species. These sequences must all originate from a common ancestor[24]. Homologous sequences of different species may differ in length, due to base pair insertions and/or deletions that occur in the course of evolution. Note, that substitution occur as well. However, they do not affect the sequences' lengths. Corresponding regions are aligned by inserting gaps ($-$) into the sequences [68]. The resulting data structure is called a *Multiple Sequence Alignment* (MSA). A MSA of DNA data is a matrix with rows corresponding to the per-taxon sequences. The columns of this matrix are called *sites*. A single site contains the presumably homologous bases of all taxa. We denote the number of sites in an MSA by m.
MSAs may contain data other than DNA data (e.g. protein data). They only differ in the set of symbols Σ, the matrix contains.

1.3.2. Categorical Data

In the scope of this thesis, we mainly work on MSAs representing categorical data. This data type occurs in a wide variety of contexts, including biology and linguistics. First, we provide a general definition of categorical data. Then, we introduce the concrete data types we base our analyses on. We further explain, how we obtain MSAs for categorical data sets. This is the basis for applying phylogenetic inference methods developed for sequence data to categorical data.

Independent of the concrete context, we represent a categorical data set as a matrix describing the following function:

$$\begin{aligned} M\colon (\mathcal{T}\times\mathcal{C}) &\to \mathcal{V}^* \\ (t,c) &\mapsto V\subset\mathcal{V}_c \text{ where } \kappa_c = |\mathcal{V}_c| \end{aligned}$$

where $\mathcal{T}$ denotes the taxon set (i.e., languages in linguistics and species in biology) and $\mathcal{C}$ the set of examined characteristics. In M, a set of values is assigned to each taxon t / characteristic c pair. The possible values $\mathcal{V}_c$ depend on the characteristic c. $\mathcal{V}$ is the union of the sets $\mathcal{V}_c\ \forall c\in\mathcal{C}$. We say that M is a *single-state matrix*, if $|M(t,c)|\leq 1\ \forall(t,c)\in\mathcal{T}\times\mathcal{C}$, otherwise, we call M a *multi-state matrix*.

In the following, we introduce specific types of categorical data in biology and linguistics. For each type, we provide the domains $\mathcal{T}$, $\mathcal{C}$, and $\mathcal{V}$.

1.3.3. Categorical Data in Biology

In biology, the taxa in $\mathcal{T}$ are the species, for which we aim to infer, for instance, a phylogeny. The characteristics in $\mathcal{C}$ are morphological traits, which biologists measure by observing specimens [16]. For a characteristic c, $\mathcal{V}_c$ can take several values, if c is the color of a flower. Sticking with this example, it is obvious, that the corresponding matrix of categorical data can contain multi-valued entries. Some characteristics are however only concerned with the presence or absence of a particular feature. In this case, $\mathcal{V}_c$ contains the two respective values only.

1.3.4. Categorical Data in Linguistics

In linguistics, each taxon in $\mathcal{T}$ corresponds to a language. We distinguish between three types of linguistic data, differing in what is considered being a characteristic and in the domain from which the values originate.

1.3.4.1. Cognate Data

Cognate data is based on changes in the vocabulary of the considered languages. When collecting this type of data, linguists work with lists of concepts or meanings, such as the Swadesh List [73]. Examining a language, linguists collect everyday words describing these concepts [21]. This results in a matrix M_{cog}, where a set of words is assigned to each language / concept pair. We regard cognate data as categorical data, with the set of characteristics $\mathcal{C}$ corresponding to the considered concepts. However, the data provided in M_{cog} is initially not categorical. It contains the individual words in the respective languages, which are not grouped into categories. To obtain a matrix M of categorical data, linguists replace the words by their *cognate classes* [21]. Cognate classes unite words, admitting a common ancestor [21]. Hence, in case of cognate data, the value set $\mathcal{V}$ contains cognate classes. In every language, there can exist words from multiple cognate classes for one concept. M is hence a multi-state matrix [21]. We assume, that linguists design concept lists in a way, that there exists at least one word in every language. We interpret it as missing information if there is no word given in M_{cog}.

1.3.4.2. Sound Class Data

The aim of sound class data is to capture changes in sound accompanying vocabulary changes [42]. If two languages admit words of different cognate classes for a certain concept, these words also differ in their sound. However, the sounds can also be different for words belonging to the same cognate class. Hence, sound class data is more fine-grained in this respect.
When considering sound class data as categorical data, the set of characteristics $\mathcal{C}$ again corresponds to the examined concepts. Each value in $\mathcal{V}$ is a tuple of a cognate class and of a sound class taken from a phonetic alphabet (e.g. as introduced by *Brown et al.* [12]). To determine $M(t, c)$, we again consider all words describing the corresponding concept in the respective language. For each word and for each sound occurring in this word, $M(t, c)$ contains the tuple of the word's cognate class and of the corresponding sound class. Thus, the resulting MSA M is a multi-state matrix.

1.3.4.3. Morphological Data

While the previously introduced data types focus on words, morphological data is based on structural features of the languages under study [20]. For collecting data, linguists evaluate the properties of the languages regarding these features. For example, they examine the number of cases in a language or whether the verb is always at the second position of a sentence. The resulting categorical matrix is structured in the same way as for morphological data in biology. We further distinguish between *morphosyntactic* and *morphophonological* data, depending on whether it contains grammatical or phonological features. Studying morphological data is of interest, because it potentially enables going further back in time and studying languages for which no written record exists [23].

1.3.5. MSA Construction

Within this section, we explain, how MSAs can be constructed for categorical data. For a set of categorical data, we are able to obtain a binary MSA A containing the symbols $\Sigma = (0, 1)$ only. For categorical data sets with a single-state matrix, we can additionally construct a multi-valued MSA A^*. It can contain up to 64 different symbols (more symbols are not possible due to technical limitations of RAxML-NG [44]) provided in an ordered list Σ_{multi}. For categorical data with a multi-state matrix, this is however only possible under restrictions.

Given a matrix M with categorical data, a binary MSA A can be obtained as the corresponding presence-absence-matrix. Each characteristic c in M is therefore represented by κ_c sites in A, each corresponding to a value v in $\mathcal{V}_c$. If $v \in M(t, c)$ for a certain taxon t, the respective entry is set to 1, otherwise to 0. Special consideration is required for the case, that $M(t, c) = \emptyset$ for some $(t, c) \in \mathcal{T} \times \mathcal{C}$. This can semantically be interpreted in two different ways: none of the considered values of c is present in t, or for each characteristic, there must be at least one value present in every taxon. In this case, an empty set indicates missing information. In the following, we use the latter interpretation. If $M(t, c) = \emptyset$, we hence set all κ_c sites to $-$.
Next, we explain, how to obtain a multi-valued alignment A^* representing M. We first examine the transformation when M is a single-state matrix. In contrast to A, A^* admits only one site for characteristic c. To determine the symbols in this site, we order the set of values in $\mathcal{V}_c$. We obtain a vector $(v[1], \ldots, v[\kappa_c])$. If $M(t, c) = \{v[i]\}$ for a taxon t, we set the respective entry to $\Sigma_{\text{multi}}[i]$, that is to the i-th symbol for multi-valued MSAs. If $M(t, c) = \emptyset$ we again assume, that data is missing, and we subsequently set the entry to $-$. To represent c in A^*, we require κ_v symbols. In the entire MSA $s_{\max} \coloneqq \max(\{\kappa_c \colon c \in \mathcal{C}\})$ different symbols occur. We note that the specific list of symbols Σ occurring in A^* is the prefix of length $s_{\max}$ of Σ_{multi}.

Constructing a multi-valued MSA for a multi-state matrix, would require an exponential number of symbols. Phylogenetic inference on such an MSA is neither technically feasible nor meaningful. However, it is possible to obtain a multi-valued MSA, which contains at least a part of the information of M. For this purpose, we discard excess elements from entries in the matrix in order to transform it into a single-state matrix. For each pair (t, c) $t \in \mathcal{T}$, $c \in \mathcal{C}$ we only keep that value in $M(t, c)$, which occurs most frequently for c over all taxa in $\mathcal{T}$. For each characteristic, we determine the ratio of taxa, for which we discard at least one value. If this ratio exceeds a fixed threshold h, we entirely discard c in order to prevent distortion in downstream analyses. Like this, we obtain a single-state matrix for which we can retrieve a multi-valued MSA containing a part of the information from M. However, we only consider this MSA in subsequent analyses, if the ratio of discarded characteristics does not exceed a second threshold g.

Example illustrating MSA construction for categorical data

Matrix M with categorical data (multi-state):

	C1	C2	C3	C4
Taxon1	$\{C\}$	$\{B\}$	$\{A\}$	$\{A, B\}$
Taxon2	$\{B\}$	$\{\}$	$\{A, B\}$	$\{A, B\}$
Taxon3	$\{A\}$	$\{A\}$	$\{C\}$	$\{A\}$

Binary MSA A:

	C1			C2		C3			C4	
	A	B	C	A	B	A	B	C	A	B
Taxon1	0	0	1	0	1	1	0	0	1	1
Taxon1	0	1	0	–	–	1	1	0	1	1
Taxon1	1	0	0	1	0	0	0	1	1	0

Matrix M after discarding excess elements:

	C1	C2	C3	C4
Taxon1	$\{C\}$	$\{B\}$	$\{A\}$	$\{A\}$
Taxon2	$\{B\}$	$\{\}$	$\{A\}$	$\{A\}$
Taxon3	$\{A\}$	$\{A\}$	$\{C\}$	$\{A\}$

Multi-valued MSA A^*:

	C1	C2	C3	C4
Taxon1	2	1	0	0
Taxon2	1	–	0	0
Taxon3	0	0	2	0

1.4. Phylogenetic Inference

Provided an MSA with data for a set of taxa, we can infer a phylogenetic tree with the methods we introduce in this section. We group these methods into distance-based and character-based methods [68]. Algorithms belonging to the first group work based on a pair-wise distance matrix for the sequences in the MSA. They tend to be faster but less accurate than the character-based methods, which infer a tree directly from the MSA data itself [68]. In the scope of this thesis, we use neighbor joining, a distance-based method, as well as Parsimony and Maximum Likelihood (ML), which are character-based. We note that ML is generally slower but more accurate than Parsimony [27, 28].

1.4.1. Neighbor Joining

Neighbor joining (NJ) is a distance-based method for phylogenetic inference introduced by *Saitou and Nei* [65]. NJ starts from a star-like tree consisting of leaves for all taxa, directly placed below a root node. First, the algorithm calculates a distance matrix containing the pairwise distances of the sequences provided for the examined taxa. NJ chooses the pair with the lowest distance and places the respective nodes below a newly created inner node. Subsequently, the algorithm updates the distance matrix by removing the entries corresponding to the selected pair and inserting the distances to the newly created node instead. NJ continues with selecting the next pair and proceeds like this until only one entry remains in the distance matrix and yields the final phylogenetic tree.

1.4.2. Parsimony

Another approach for the reconstruction of phylogenetic trees was introduced by *Farris* [26] and *Fitch* [31]. It is based on the *parsimony criterion*, which favors trees requiring a minimum number of substitutions to explain the sequences. Given an MSA and a respective phylogenetic tree, the parsimony score is obtained as the sum over the scores calculated separately for each MSA site. The per-site parsimony score of a given tree is the minimum number of substitutions which must occur over the entire tree to generate the observed data at the leaves.
A parsimony heuristic performs a tree search in order to return a tree minimizing the parsimony score. Note, that this algorithm is in general not deterministic, as there can exist several trees yielding the same parsimony score. Further, it is not guaranteed, that a tree with a minimum parsimony score is returned, as the problem is $\mathcal{NP}$-hard [30] and the tree search is only a heuristic.

1.4.3. Maximum Likelihood

Maximum likelihood (ML) [29] is another method for phylogenetic inference that involves finding a tree optimizing a function. Instead of minimizing the number of necessary substitutions, however, the goal is to maximize the likelihood $L(\Theta|D)$ for the given MSA D. The parameter vector $\Theta \coloneqq (T, b, \mathcal{M}, \phi)$ comprises the topology T of the tree, the vector b containing its branch lengths, a substitution model $\mathcal{M}$ (see Section 1.4.3.1) and a set of internal parameters which we denote by ϕ. We define the corresponding likelihood as $L(\Theta|D) \coloneqq P(D|\Theta)$. It hence corresponds to the probability of the given MSA D to arise under the setting described by the parameter vector Θ. In other words, ML tries to find the tree best explaining the observed data.

1.4.3.1. Substitution Models

Let D be an MSA containing $s_{\max}$ symbols $\Sigma = (\Sigma[1], \ldots, \Sigma[s_{\max}])$. We define a *substitution model* $\mathcal{M}$ as a tuple (π, R) [80]. For $i \in (1, \ldots, s_{\max})$, $\pi[i]$ is the probability with which $\Sigma[i]$ initially occurs. We denote the probabilities in π as *equilibrium frequencies*. As they sum up to 1, $\pi[s_{\max}]$ can be given relatively to the remaining entries. The vector is hence determined by $s_{\max} - 1$ values.
For each $i, j \in (1, \ldots, s_{\max})$, $R[i, j]$ provides the rate of substitution between $\Sigma[i]$ and $\Sigma[j]$. As we model evolution as a continuous time Markov Chain, this value depends on nothing but $\Sigma[i]$. Further, we assume that evolution is a time reversible process. Hence, it holds that $\pi[i]R[i, j] = \Pi[j]R[j, i]$ $\forall i, j \in (1, \ldots, s_{\max})$. We further note that the values of each row of R must sum to 0. R is therefore defined by the $\frac{1}{2}s_{\max}(s_{\max} - 1)$ values in the strict lower triangle matrix. Additionally, R is normalized. Subsequently, one value less is required for unambiguous determination. The whole substitution model hence comprises $s_{\max} + \frac{1}{2}s_{\max}(s_{\max} - 1) - 2$ parameters, where $s_{\max} - 1$ defines π and $\frac{1}{2}s_{\max}(s_{\max} - 1) - 1$ defines R.

Using the General Time Reversible (GTR) model [74], all parameters defining $\mathcal{M}$ are free parameters which are estimated independently from the data during the tree inference. We note that the number of free parameters of this substitution model is quadratic with $s_{\max}$. This can lead to overparameterization and increased runtimes. An alternative to this, that alleviates both potential problems, is the MK model [47]. In this model, it holds that $\Pi[i] = \frac{1}{s_{\max}}$ $\forall s[i] \in \Sigma$ and $R[i][j] = 1.0$ for $\forall i, j \in (1, \ldots, s_{\max}), i \neq j$. Hence, MK does not admit any free parameter. Working with a binary MSA, we henceforth only use the GTR model in our analyses, which we denote by BIN. As there are two different symbols in binary MSAs, GTR only has one free parameter, one equilibrium frequency. This prevents the aforementioned overparameterization and increased runtimes. For binary MSAs, GTR and MK differ only regarding the stationary frequencies, so that we omit a separate analysis under MK.

Further, we allow it, to extend a substitution model in order to incorporate *rate heterogeneity* among sites to accommodate that sites of an MSA evolve under different rates [79]. For this purpose, we multiply the rate matrix for the k-th site with a factor r_k. We assume that these factors are Γ-distributed. Modelling rate heterogeneity with this approach leads to an additional free parameter α, which determines the shape of this Γ-distribution. Since its introduction in the mid 1990s by *Yang* [79], the Γ-model is widely used approach for modelling rate heterogeneity. We add +G to a model's name if we aim to indicate, that we additionally use the Γ-model.

1.4.3.2. Likelihood Computation

Based on [80], we describe how to compute the likelihood under a given MSA D for $\Theta = (T, b, \mathcal{M}, \phi)$ and, in particular, for a tree with topology T and branch lengths b. Assuming that the sites of D evolve independently, we obtain this score as the product over the per-site likelihoods under Θ. The per-site likelihoods are usually very small. In practice, to avoid underflow, we therefore calculate the per-site *log*-likelihoods, and the final *log*-likelihood is the sum over all per-site log-likelihoods of the MSA.

In order to explain the computation of the per-site likelihood for a fixed site, we first assume, that we are given the inner states, that is, symbols observed for this site at the inner nodes of the tree. For each branch with incident nodes n_1 and n_2, we determine the probability, that the symbol observed at n_1 is substituted by the symbol observed at n_2 over the time represented by the length of this branch. This probability depends on the rate matrix R of the model $\mathcal{M}$. We calculate the product over the probabilities for all branches. Additionally, we multiply the result with the equilibrium frequency of the symbol observed at the root. Like this, we obtain the per-site likelihood for a tree with given inner states. During the inference, however, only trees with unknown inner states occur. To calculate the per-site likelihood for such a tree, we sum the per-site likelihoods over all possible trees with fixed inner states. We use the Felsenstein Pruning Algorithm [27], which efficiently calculates the likelihood over all possible combinations of inner states via dynamic programming.

1.4.3.3. Likelihood Optimization

Finding a tree with a maximum likelihood score requires evaluating the likelihood of all possible trees. As the number of existing tree topologies grows super-exponentially with the number of taxa, this results in super-exponential runtime. The optimization problem is hence $\mathcal{NP}$-hard [15]. We infer trees with the help of RAxML-NG [44] which uses the heuristic introduced in the following. We start from an initial assignment of the parameter vector Θ. This requires a topology, which we either obtain via parsimony (see Section 1.4.2) or we use a random tree. In the following, we iteratively optimize Θ to improve the likelihood. The possible adaptions are:

- Change the tree topology T with the help of topological moves [80]
- Numerical optimizations of model parameters via Brent's method [11] or the Broyden–Fletcher–Goldfarb–Shanno method [32]
- Optimizations of branch lengths with the Newton-Raphson method [44]

1.4.3.4. Quantifying Difficulty

The heuristic for the inference of ML trees does not guarantee that we will find a tree with the global maximum likelihood score. Instead, it is possible, that the search algorithm converges to a local maximum of the likelihood distribution over the tree space. We usually perform several independent tree inferences. If they return similar trees (i.e., topologies that admit low RF distances to each other) this indicates that the likelihood distribution has a clear peak, and we refer to the respective MSA as being easy to analyze. If multiple tree inferences result in different topologies with high RF distances, the likelihood surface is more rugged, and we observe multiple local maxima. We classify the corresponding MSA as being more difficult.
A score for quantifying the difficulty of a phylogenetic inference on a given MSA was introduced by *Haag et al.* [39]. It is based on 100 tree inferences with RAxML-NG. Among the ML trees resulting from these inferences, we determine the best tree, that is the one with the highest log-likelihood score. We apply all statistical significance tests implemented in IQ-TREE [56] to these ML trees. Those trees, which are not significantly wore than the best tree, are deemed *plausible* [54]. If there are more plausible trees among the ML trees, the introduced difficulty score is lower. The score increases with the mean RF distance among all ML trees and the mean RF distance among the plausible trees. Further, it leads to a higher difficulty score, if there is a higher ratio of unique topologies among all ML trees or among the plausible trees.

1.5. Contribution

When August Schleicher got in touch with Darwin's book "On the Origin of Species", he recognized major parallels between the evolution of species and languages. He described them in his letter "The Darwinian Theory and the Science of Language" [66] and proposed to use scientific methods from evolutionary biology in historical linguistics. In the following we discuss possible challenges of applying methods from computational biology to linguistics. *Ladoukakis et al.* [45] propose a mapping of evolutionary concepts in linguistics and biology, showing that this is only possible to a limited extent. In biology, one can distinguish between the genetic material of an individual and its outer appearance. For languages, there is no comparable subdivision. Characteristics can be easily transferred between languages through contact between speakers [61], resulting in a high geographical bias in the data. Different components of languages (lexicon, grammar, syntax) may further evolve differently. As a consequence, it is not possible, to infer a single ground truth phylogeny for a set of languages [38].
Moreover, we must not neglect the steps preceding the phylogenetic inference. When we construct an MSA for a categorical data set, we lose any semantic information associated with the values of a characteristic. If the values correspond to ordered numbers or admit any other hierarchical relationships, this is not captured by the MSA. In binary MSAs, the sites representing the same characteristic are negatively correlated with each other, and thus they are not necessarily independently identically distributed. However, this is what we typically implicitly assume when applying the inference methods presented. For cognate data and sound class data, an additional pre-processing step is required. We determine the words' cognate classes in order to obtain a categorical data set. This imposes the knowledge related with the classification on the data and can bias it. We might further lose information due to the grouping (e.g., the degree of similarity of words in the same class). It is also unclear, how to handle synonyms [49].
Finally, we note that biologists and linguists proceed differently when reconstructing phylogenies. A key difference is that linguists often have reference trees available, which they have constructed manually [40]. On the one hand, this offers the chance to assess the result of an inference by comparing it to the reference. On the other hand, it poses the risk of biasing the inference, for instance by tweaking parameters and data until a desired result is achieved.

Although languages and species both form evolvable systems, there are hence substantial differences between the evolutionary processes in the two domains. Applying methods for phylogenetic inference is therefore related with numerous challenges, and this thesis represents only a first step towards facing them.
In Chapter 2, we conduct a case study on a morpho-syntactic MSA, exploring several approaches for accommodating the geographical bias in the data. In Chapter 3, we analyze a set of linguistic MSAs and compare them to biological morphological MSAs. In this context, we also contrast binary and multi-valued MSAs. This provides an insight into the challenge of transforming categorical data into a format that is suitable for phylogenetic inference. In Chapter 4, we address the issue of handling synonyms and we propose a potential solution.

2. Case Study on a Morphosyntactic MSA

In this chapter, we conduct a case study for a family of 46 Indo-European languages. This language family has been well studied in classical historical linguistics and except some minor debates, linguists arrived at a consensus about the evolution of these languages. Using phylogenetic methods, we are able to infer a tree T_C, which is in line with that knowledge. The respective inference is based on an MSA A_C of cognate data, derived from *Bouckaert et al.* [10]. Additionally, we have a second MSA, A available, containing the same set of languages, but the MSA is based on morphosyntactic data obtained from the WALS database [20] (see Section 2.1.1). However, it is not possible to reproduce T_C using the MSA A [53]. In this chapter, we investigate alternative approaches, to determine whether the MSA A contains signal indicative of T_C.

We suspect, that not only vertical but also horizontal and convergent evolution occurs at the sites of A. Vertical evolution refers to changes over time, whereas horizontal evolution is contact-induced due to geographical proximity. We assume that the consensus tree T_C reflects vertical evolution. Horizontal evolution supposedly occurs according to geographical proximity of areas, where the respective languages are spoken. These geographical relationships are represented by a geographical tree T_G. Languages which admit similarities that are not due to a common origin but to proximity and contact among speakers are called a Sprachbund [52]. The languages spoken in the Balkan region, for example, share common features despite their different origins [64]. Horizontal evolution also occurs during the evolution of species. In this context, it is referred to as horizontal or lateral gene transfer [62]. Convergent evolution results in homoplastic sites, which means that similar traits evolved in independent lineages [75]. This phenomenon is also common to the evolution of languages and species.

For morphosyntactic data it is not fully studied, which characteristics evolved vertically, horizontally, or convergently. As a consequence, phylogenetic methods do not only group languages on the basis of having the same origin, but also based on geographical relationships. Consequently, a Sprachbund may for example be included in the inferred phylogeny. Reconstructing the signal for T_C from A might therefore be related to classifying the sites in the MSA into vertically and horizontally evolved.

After introducing the materials we used (Section 2.1), in Section 2.2 we investigate to which extent known methods provide information about the signal contained in A. In Section 2.3, we present and evaluate an algorithm for determining sites supporting T_C, and in Section 2.4 we introduce a new mixture model for maximum likelihood computations taking geography into account. We further examine, whether this model's purpose to handle the differently evolving sites in A is fulfilled.

2.1. Material

2.1.1. The MSA

The morphosyntactic MSA we analyze is based on data we obtained from the WALS database[20]. The MSA encodes 92 characteristics (features) for 46 languages. These features cover a variety of grammatical phenomena, classified into nine categories: phonology (20), morphology (12), nominal categories (29), nominal syntax (8), verbal categories (17), word order (56), simple clauses (26), complex sentences (7), and lexicon (13) [53]. All features are multi-valued, but single-state. We enrich these features by additional information according to *Michelioudakis et al.* [53] and derive a binary MSA A as described in Section 1.3.5. This MSA consists of 425 sites, with 219 being informative, meaning that they contain at least one 1 and one 0. We denote the MSA only containing the informative sites of A as $\hat{A}$.

2.1.2. Reference Tree

Within this scope, the tree T_C is used as a reference for the evolution of the languages in the considered family. The respective cladogram is provided in Figure 2.1. T_C is constructed based on the cognate MSA A_C, applying Bayesian inference with BEAST [72]. Hence, this tree is not a result from the methods of classical historical linguistics, but its coarse structure agrees with the findings from that field.

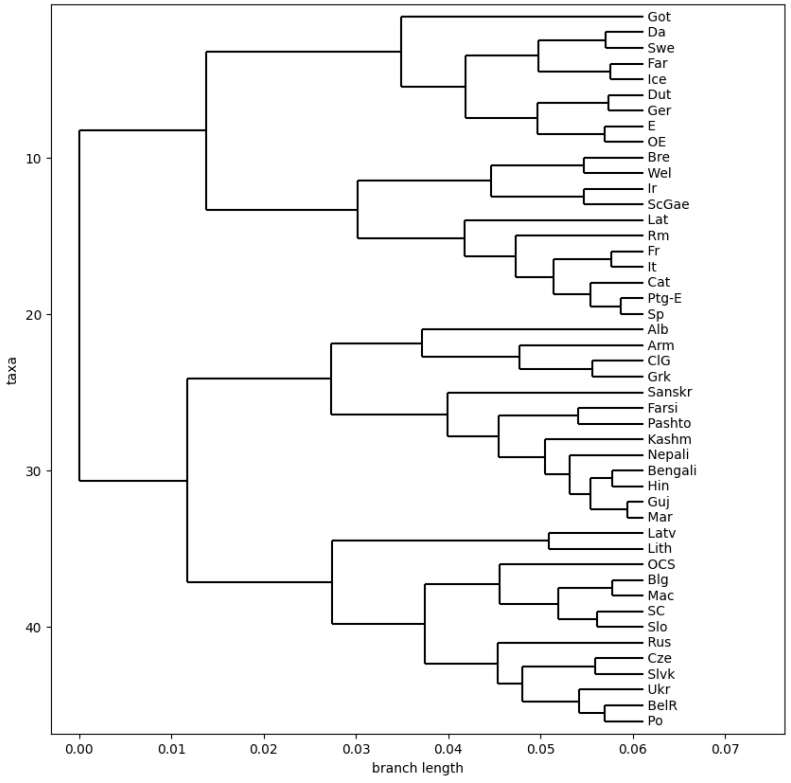


Figure 2.1.: Cladogram of T_C

2.1.3. Geographical Tree

In order to accommodate horizontal evolution, we require a geographical tree T_G representing spatial proximity.
To construct such a tree, we have to assign a location to each language. For this purpose, we use the centroids published by *Gray and Atkinson* [36]. Based on these coordinates, we determine a distance matrix, which we subsequently use to construct a tree via neighbor joining. For calculating the distances, we apply three different methods. Firstly, we consider the geodesic distances calculated with the haversine distance formula [19]. In addition, we use route planning with routingpy (`https://pypi.org/project/routingpy/`) and compute a connection route for each centroid pair. As a distance metric, we then consider both the path length as well as the temporal duration of these routes. Each distance metric yields a distinct tree (see Appendix A). In the following, we use the tree based on route duration because we consider it as most suitable.
The choice of the tree based on route duration appears to be most plausible from a semantic perspective. We aim to construct a tree capturing contact-induced transfer of language properties. For such a transfer to occur, speakers of different languages must be in contact with each other and therefore move away from areas where only or mainly their own language is being spoken. In doing so, they move along roads, which is modeled by route planning. Considering the temporal duration of routes instead of their spatial length takes into account topographic information. If one compares two routes of equal length, one traversing flat terrain and one mountainous terrain, the latter will require more time. According to this, if speakers of two languages are separated by a mountain range, they are less likely to establish contact than if there is no topographic obstacle at the same spatial distance.
The branch lengths in the resulting geographical tree correspond to duration in seconds. Thus, they are in a different range than the branch lengths of characteristic phylogenetic trees, which typically lie between 0 and 1. Using the geographical tree with the original branch lengths leads to distractions and impacts the use of phylogenetic analysis tools, for example the branch length optimization in RAxML-NG. We therefore develop different approaches for transforming the branch lengths to a more representative range. All our methods are based on the branch lengths occurring in a reference tree set. This set contains the plausible trees we obtained from 400 ML tree searches with RAxML-NG on the morphosyntactic MSA A (The experiments to generate these plausible trees are introduced in greater detail in Section 2.2.5). The branch lengths of all trees in the reference tree set range between 0.0 and 0.46, with an average of 0.03. Our first approach is to scale the branches of the geographical tree such that they have the same average length as the reference trees. Our second approach is to transform the branch lengths to the same range as branch lengths occurring in the reference tree set. With our final approach, we generate trees having the same topology as the geographical tree, but with branch lengths randomly sampled from the branch length distribution in the reference tree set.
We examine the effect of each branch length transformation on the likelihood of the geographical tree respective to the alignment A. Already without branch length optimization, all trees with transformed branch lengths admit a significantly better likelihood than the original tree. If we determine the likelihood of the trees with transformed branch lengths using branch length optimization, they all admit exactly the same likelihood. This is because the optimization results in exactly the same branch lengths for all of these trees. However, it is not possible, to obtain this result if we omit the branch length transformation. Overall, the results show that while the transformation of branch lengths is crucial, the choice of the specific method only plays a minor role. We therefore decide to scale the tree to the same average branch length as observed in the reference tree set.

The tree we refer to as T_G in the following is hence the geographical tree constructed based on route duration, scaled to the average branch length in the reference tree set. It is illustrated in Figure 2.2.
Note that the topologies of the geographical tree T_G and the consensus tree T_C are highly distinct, with an RF distance of 0.81.

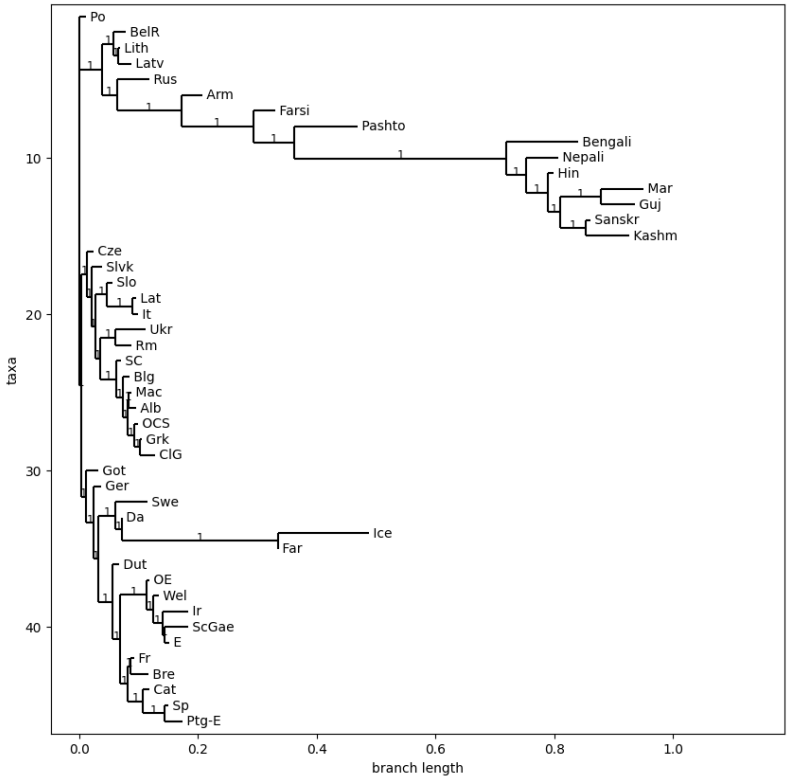


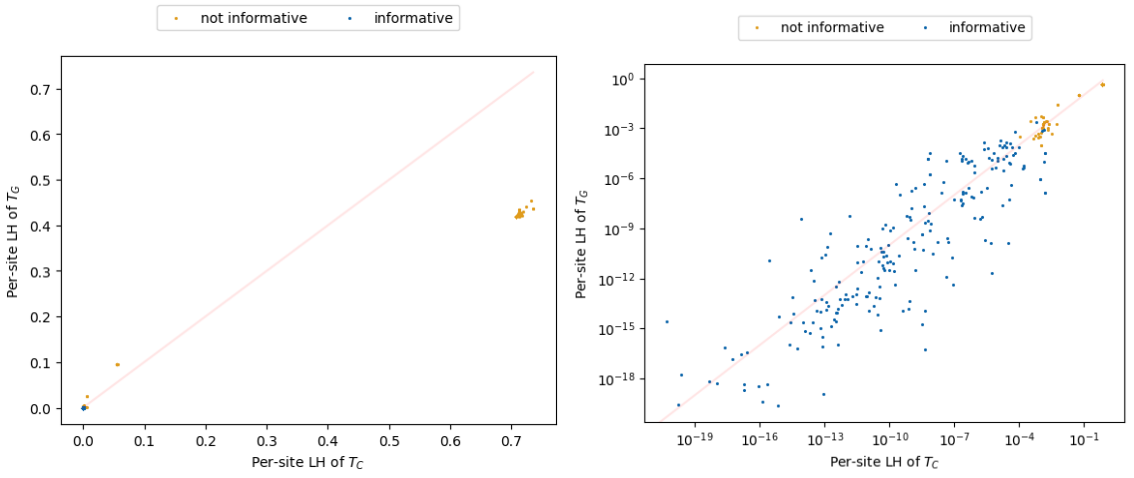
Figure 2.2.: Cladogram of T_G

2.2. Signal Recognition

In order to analyze to which degree the information provided in the sites of the MSA A is congruent with the consensus tree T_C (representing vertical evolution) and with the geographical tree T_G (representing horizontal evolution), we deploy of several methods for identifying this signal. We investigate two per-site metrics on A, namely per-site likelihood (see Section 2.2.1) and weight calibration (see Section 2.2.2). Further insights are provided from rooting the trees T_C and T_G (see Section 2.2.3). Additionally, we use the delta statistics, a metric for trait association (see Section 2.2.4), and in Section 2.2.5 we assess the results of a tree inference with RAxML-NG.

2.2.1. Per-Site Likelihood

The per-site likelihood is the likelihood of a tree for a single site of the respective MSA only. Note that values are actual likelihoods, not log-likelihoods. Higher values indicate stronger support of the respective site for the given tree. In Figure 2.3a, the per-site likelihoods with respect to T_C are shown on the x-axis, those with respect to T_G on the y-axis. Each marker corresponds to a site in A, with blue markers indicating that the site is informative, and orange ones indicating that the site is not informative. For both trees, all informative sites yield a per-site likelihood close to 0, most of the non-informative sites yield higher per-site likelihoods (approximately 0.7 in T_C, approximately 0.3 in T_G). For MSA A, the per-site likelihoods for both trees are clearly correlated, with a Pearson correlation coefficient of 1.00 (see Figure 2.3b; p-value $< 10e-8$). Considering $\hat{A}$ with informative sites only (corresponding to blue markers in the figure), the correlation between T_C and T_G is lower, with a correlation coefficient of 0.30 (p-value 0.09). We conclude, that we are not able to separate vertically and horizontally evolving sites based on per-site likelihoods.



(a) Linear scale

(b) Logarithmic scale

Figure 2.3.: Per-site likelihoods are indicated on the x-axis for T_C, on the y-axis for T_G. Each marker corresponds to a site in A, blue markers indicate informative sites, orange markers indicate non-informative sites. The values for both trees are clearly correlated. All informative sites admit per-site likelihoods close to 0.

2.2.2. Weight Calibration

An alternative per-site metric is weight calibration. To calculate it, we generate 100 random trees and determine the per-site likelihoods for each of these trees. The weight calibration of a site corresponds to the number of random trees where its per-site likelihood is worse than in the reference tree [7]. Hence, the values range from 0 to 100. The higher the per-site weight, the more clearly the site is congruent with the tree. Figure 2.4 depicts the weight calibration of the sites in A with respect to the consensus tree and the geographical tree. In the consensus tree T_C, most sites show either a weight calibration of almost 100 or below 20. We observe similar results for the geographical tree T_G, although the separation is less pronounced. The correlation of weight calibration values between both trees is lower than for the per-site likelihoods, with a Pearson correlation coefficient for all sites of 0.76 (p-value 0.04). This is caused by sites, which admit a high weight calibration with respect to one tree but a low value with respect to the other tree. However, only few such sites are additionally informative. The Pearson correlation coefficient considering informative sites only is 0.71 (p-value 0.03). We conclude that we cannot distinguish between vertically and horizontally evolving sites based on weight calibration.

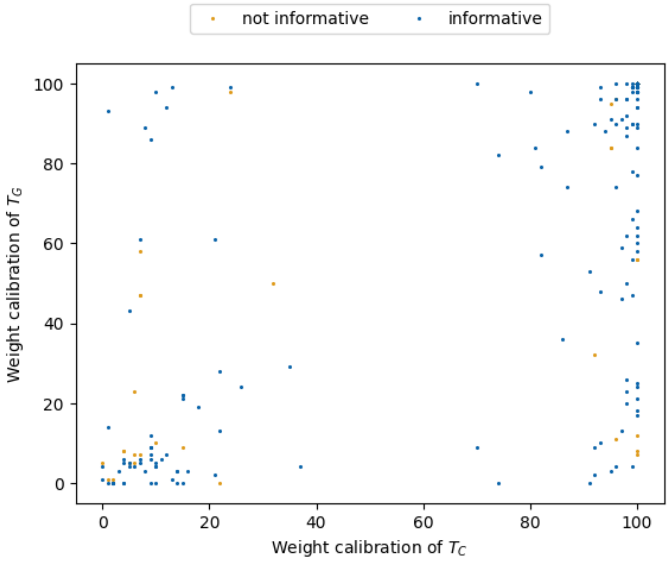
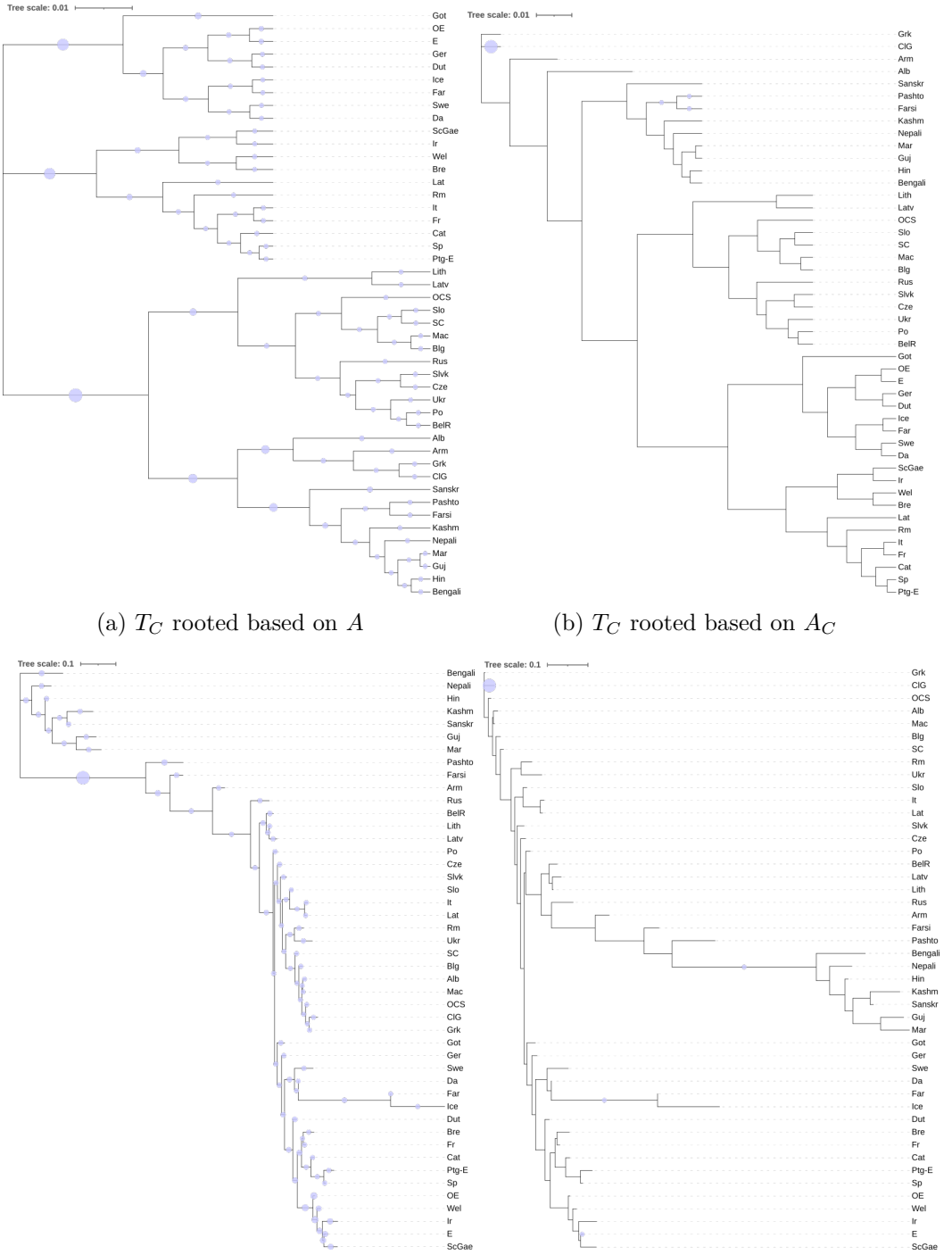


Figure 2.4.: Weight calibration is indicated on the x-axis for T_C, on the y-axis for T_G. Each marker corresponds to a site in A, informative sites are colored in blue, non-informative sites in orange.

2.2.3. Rooting

In the following, we present observations related to rooting the trees T_C and T_G. This was primarily motivated by the fact that investigating trait association (see Section 2.2.4) requires rooted versions of T_C and T_G. However, the application of Root Digger [8] to the trees and MSAs considered provided some interesting insights on its own. The results do not allow for immediate conclusions about which sides of A evolve horizontally and vertically, but still provide information about the phylogenetic signal contained in the MSAs A and A_C.
Root Digger leverages ML with an non-reversible model to infer a root and requires an MSA in addition to the tree. Concerning T_C it is reasonable, to use the cognate MSA A_C, as T_C is inferred based on A_C (see Section 2.1.2). Nevertheless, we also investigate, the behavior of Root Digger, when we use MSA A instead. We constructed T_G without an MSA (see Section 2.1.3). Hence, we consider both A and A_C as an input MSA for Root Digger, and assess, how the results differ. Figure 2.5 illustrates the rooted trees.
To evaluate the confidence of the root placement as determined by Root Digger, we make use of likelihood weight ratios (LWR) as introduced by *Strimmer and Rambaut* [71]. Let L_i be the likelihood we obtain, when choosing a node i as root of the tree. The LWR of i is then determined as $L_i / \sum_{j \neq i} L_j$. The LWRs of all nodes sum up to 1. A high LWR suggests choosing the respective node as the root.

First, we analyze the LWR distributions, we obtain, when using the MSA A_C as an input for Root Digger. For both T_C and T_G, the node, which is selected as the root yields an LWR $> 1 - (10\text{e} - 4)$. All other LWRs are comparatively close to 0. The choice of the root is hence supported by a clear signal in the MSA.
If we do the same calculations using the morphosyntactic alignment A, the root determined in T_C yields an LWR of 0.16 only. As the LWRs sum up to one, this low maximum value implies a more uniform (or fuzzy) distribution of the LWRs over the branches of the tree. As a consequence, we are substantially less confident regarding the root placement. In T_G, the same effect is even more pronounced, with a maximum LWR of only 0.09 for the node returned as a root. The fact that it is not possible to determine a confident root in both trees indicates an insufficient signal in the morphosyntactic MSA A.



(a) T_C rooted based on A (b) T_C rooted based on A_C

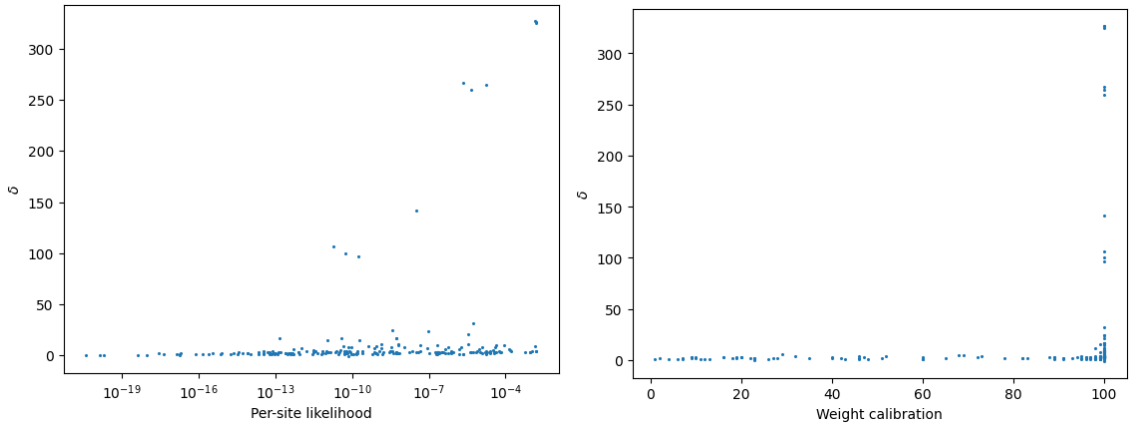
(c) T_G rooted based on A (d) T_G rooted based on A_C

Figure 2.5.: Trees rooted with Root Digger.
The size of the circles on the branches corresponds to the LWR score for rooting the tree at the respective position. If there is no circle, this corresponds to LWR = 0

2.2.4. Trait Association

In this section, we consider an approach, which is not based on a per-site metric, but can still be used to capture the signal contained in the sites of A. Given a phylogenetic tree and a trait of the involved species, the phylogenetic trait association (also referred to as phylogenetic signal) describes the degree to which species related in the tree tend to admit the same trait characteristic [81]. We aim to identify vertically and horizontally evolving sites in the MSA A. This problem can be regarded as a trait association question with respect to T_C and T_G, respectively, with each site of the MSA being considered as an individual trait.
Researchers developed several metrics to measure trait association [34, 1, 58]. For this case study, we exclusively use the δ statistics introduced by *Borges et al.* [9]. As this metric requires a rooted tree, we make use of the roots we obtained for T_C and T_G using Root Digger and the cognate MSA A_C (see Section 2.2.3).
Applying the δ statistics to the phylogenetic tree and to a trait of its taxa yields a value δ. This value measures the entropy of the ancestral states in the tree with respect to the trait under consideration. A higher δ implies a higher degree of trait association. This means, the considered trait is more related to the proximity of the taxa as implied by the tree.

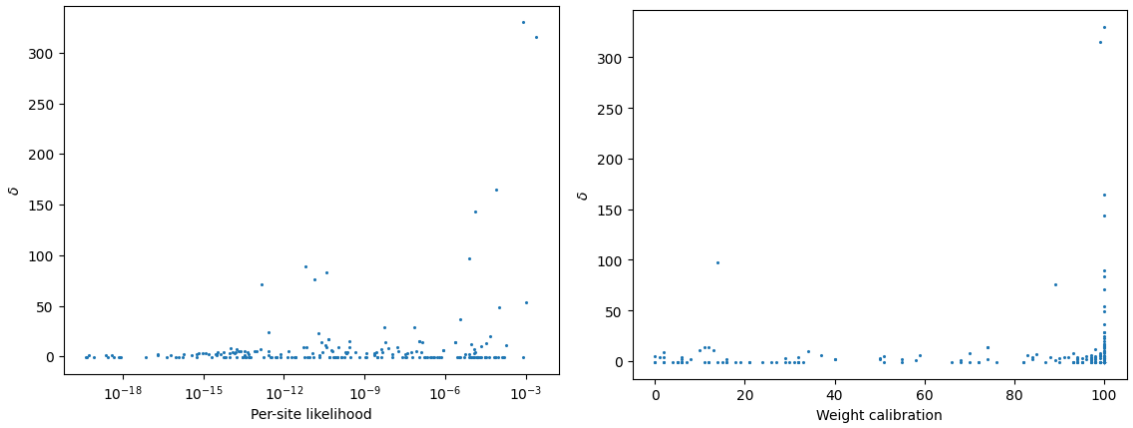
We treat each site of the informative subaligment $\hat{A}$ as a binary trait, and we determine δ with respect to T_C. The results are illustrated in Figure 2.6a and Figure 2.6b. In both figures, the y-coordinate of each marker corresponds to the value of δ for the respective site. The x-axis depicts per-site likelihoods in Figure 2.6a and weight calibration values in Figure 2.6b. We observe that only few sites exist which admit a high δ value. All such sites also have a high weight calibration value and tend to admit a higher per-site likelihood.



(a) Correlation with per-site likelihood (b) Correlation with weight calibration

Figure 2.6.: Evaluation of delta statistics with respect to T_C. Each marker corresponds to a site of $\hat{A}$. In both plots, the y-axis represents δ regarding T_C. In Figure 2.6a, the x-axis indicates per-site likelihoods, in Figure 2.6b, it indicates weight calibration.

Further, we determined δ for each site of $\hat{A}$ with respect to T_G instead of T_C. We observe the same tendency as for the consensus tree T_C (see Figure 2.7). However, the δ values tend to be lower and the correlation to the other metrics is less pronounced.



(a) Correlation with per-site likelihood (b) Correlation with weight calibration

Figure 2.7.: Evaluation of delta statistics with respect to T_G. Each marker corresponds to a site of $\hat{A}$. In both plots, the y-axis represents δ with respect to T_G. In Figure 2.7a, the x-axis indicates per-site likelihoods, in Figure 2.7b, it indicates weight calibration.

Given the δ values with respect to T_C and T_G, we subsequently analyze the relationship between the two distributions. In Figure 2.8, δ regarding T_C is indicated on the x-axis, δ regarding T_G on the y-axis. We find no correlation for the respective values of both trees (Pearson correlation coefficient 0.08; p-value 0.25). However, the proportion of sites with a clear phylogenetic signal is low in both trees.
If we want to identify a vertically evolving site based on the respective trait association, we expect it to yield a high δ value regarding T_C and a low δ value regarding T_G. For horizontally evolving sites, it should be the other way around. There are only 6 sites, for which we observe $\delta > 200$ regarding T_C and $\delta < 50$ regarding. The other way around, there are even only 2 sites. Thus, for the vast majority of sites in the MSA A, it is not possible to conclude from the results of the δ statistics whether they are evolving vertically or horizontally.

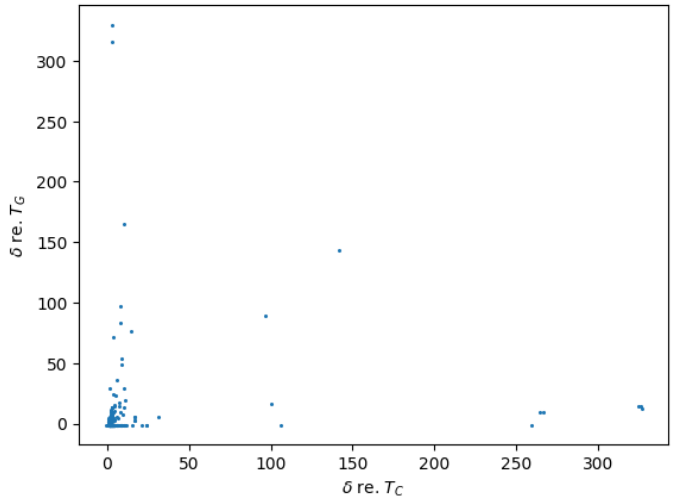
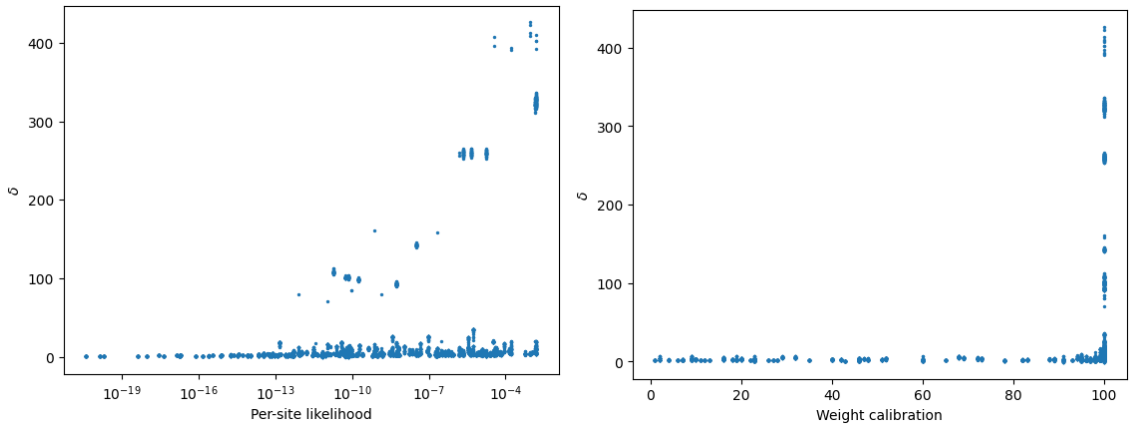


Figure 2.8.: Correlation of δ values for T_C and T_G. The x-axis represents δ with respect to T_C, the y-axis δ with respect to T_G.



(a) Correlation with per-site likelihood

(b) Correlation with weight calibration

Figure 2.9.: Evaluation of delta statistics on a set of trees, obtained from T_C by selecting each inner node as a root. Each marker corresponds to one tree and one site of $\hat{A}$. In both plots, the y-axis represents δ regarding T_C. In Figure 2.6a, the x-axis indicates per-site likelihoods, in Figure 2.6b, it indicates weight calibration.

Finally, we investigate the impact of the position of the root on the results of the δ statistics. For this purpose, we re-root T_C at every inner node, resulting in a set of trees that differ only in the root position. For each of those differently rooted trees, we calculate δ for all sites of $\hat{A}$. Figure 2.9a depicts the results of this extended analysis. Note that the per-site likelihood and the weight calibration for a site remain the same regardless of the root placement. Regarding δ, we only observe only slight fluctuations compared to the evaluation on a single root. To quantify these fluctuations, we consider the distribution of δ for each site regarding different positions of the root. We determine the coefficient of variation, which is defined as the standard deviation divided by the mean [25]. For 206 of 216 sites, we obtain a coefficient of variation ≤ 1. The δ statistics is therefore in general independent of the position of the root on our data.

2.2.5. Maximum Likelihood Tree Searches

In order to gain a depper understanding of the properties and the structure of the MSA A, we conduct ML inferences with RAxML-NG. We run 200 tree inferences, starting from 100 random and 100 parsimony starting trees. We use both BIN and BIN+G as model of evolution in two separate settings. We restrict the analysis to A as we gain no further insights by running the same experiments for $\hat{A}$, as $\hat{A}$ only lacks the non-informative sites. We can omit these sites since they do not contain a signal for a specific tree topology and thus only have a minor impact on the results of the ML computations.
First, we analyze the general convergence behavior of RAxML-NG on A. We run 200 tree inferences independently, with each tree inference returning the tree with the highest encountered ML score. This results in 200 maximum likelihood trees (ML trees) per model. Using BIN as a model, the average RF distance between all trees in this tree set is 0.21. There are 24 unique topologies among the ML trees, 21 of them are considered as being plausible (see Section 1.4.3.4). Using the BIN+G model, the ML trees admit an average RF-Distance of 0.13 with 17 distinct topologies and 13 plausible tree topologies. These results indicate a clear convergence behavior, meaning that the tree searches tend to converge to the same peak of the ML distribution. Additionally, we analyze the trees RAxML-NG generates as intermediate results during the tree searches. Note that this set of trees also contains all final ML trees. If we use BIN as a model, we observe 34 unique topologies among the intermediate trees, with BIN+G we only encounter 28 unique topologies during the tree inferences. The relatively small number of unique topologies compared to the large number of executed tree inferences confirms the observed clear convergence behaviour.

In the following, we aim to examine, whether convergence occurs towards T_C or T_G. The ML trees returned from the tree searches with BIN as a model exhibit an average RF distance of 0.82 to the consensus tree T_C. For trees inferred under BIN+G the average RF distance to T_C is 0.85. Concerning T_G, the average RF distance of the ML trees is 0.94 under BIN and 0.93 under BIN+G, respectively. According to this high distances, the trees retrieved with the help of ML inferences differ clearly from both reference trees. Using this method, it is thus not possible to recover any signal from the alignment, which would support either T_C or T_G.

2.3. Pavlidis Algorithm

In this section, we present an approach, which aims to identify a subset of vertically evolving sites of A, that is, sites evolving along the consensus tree T_C. On the resulting subalignment we conduct the same analysis as described in Section 2.2.5 and we analyze, whether the selection of sites improves the convergence of the tree inferences towards T_C.

Our work is based on an algorithm proposed by Pavlos Pavlidis. We only consider sites of MSA $\hat{A}$, meaning sites that are non-informative are immediately discarded. The proposed algorithm works iteratively. We initialize the algorithm with a subalignment of 150 randomly chosen sites of $\hat{A}$. In each round, we infer a tree for the current subalignment and compute its RF distance to the consensus tree T_C. If the current tree is closer to T_C than any tree inferred in a previous iteration, we store the current subalignment. Next, we construct a new subalignment based on the current one by replacing a small proportion of sites by randomly chosen sites of $\hat{A}$. We determine the probability of replacement or addition of a site such that the expected size of the subalignment remains constant. Like this, the subalignment only changes slightly in every step. This is important for the convergence of the algorithm.
We assume that the final subalignment resulting from the algorithm only contains the vertically evolving sites, while horizontally evolving sites have been eliminated. In the original design, Pavlidis applied neighbor joining (Alg_{NJ}) to infer a tree for each subalignment. We further tested using maximum parsimony (Alg_{Pars}) or maximum likelihood (Alg_{ML}) instead. We run Alg_{NJ} and Alg_{Pars} for 10 000 000 iterations. For technical reasons, we stop Alg_{ML} after 4 210 000 iterations. For the following evaluation, after every 10 000th iteration we sample the tree with the minimum RF distance encountered so far and the respective subalignment. We denote the last sampled subalignment by A_{NJ}, A_{Pars}, and A_{ML} respectively. The trees inferred on the base of these subalignments admit a minimum RF distance to T_C compared to all trees encountered with the respective version of the algorithm.

2.3.1. Convergence Behavior

In our first analysis, we investigate the convergence behavior of the three versions of the algorithm and analyze the resulting MSAs A_{NJ}, A_{Pars} and A_{ML}. Figure 2.10 depicts the size of the subalignment (y-axis) per iteration (x-axis). Figure 2.11 depicts the RF distance of the inferred tree to the consensus tree (y-axis) per iteration (x-axis). Each color corresponds to a distinct version of the algorithm (Alg_{NJ}, Alg_{Pars}, and Alg_{ML}).

With Alg_{NJ} we observe a slow convergence behavior, leading to a relatively low minimum RF distance of 0.14, which occurs the first time within the final 100 000 iterations executed. The resulting subalignment A_{NJ} contains 43 sites, and thus only around 10% of the 425 sites in the original MSA. With Alg_{Pars} we observe convergence to a comparatively high RF distance of 0.47. Figure 2.11 shows, that a tree exhibiting this distance is inferred substantially earlier than with Alg_{NJ}. Moreover, we observe, that the subset of chosen sites is more stable in Alg_{Pars}. With a length of 113, A_{Pars} contains more than twice as many sites as A_{NJ}. We are not able to observe a connection between the weight calibration of a site and its selection by the Pavlidis Algorithm. With Alg_{ML}, we are only able to run the first 4 210 000 iterations. The minimum RF distance we observe at this point is 0.33. A corresponding tree occurs about as early as in Alg_{Pars} (see Figure 2.11), but with Alg_{ML}, we observe more fluctuation regarding the number of selected sites (see Figure 2.10). Considering the ML trees inferred on the respective subalignments, we however only observe a minor effect of these changes in which sites are selected. A_{ML} admits 44 sites, approximately as many as A_{NJ}.

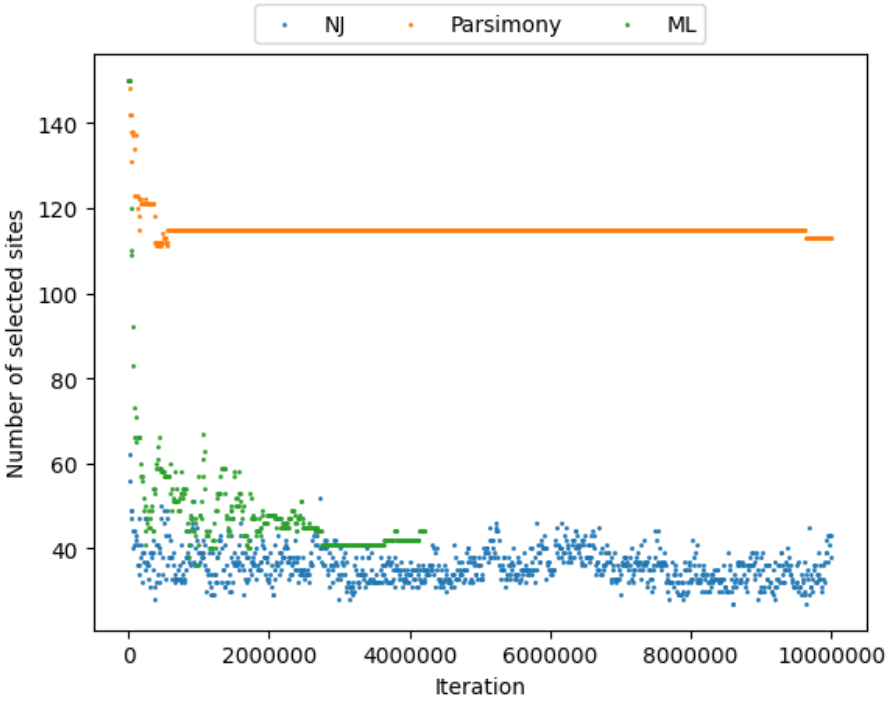


Figure 2.10.: Number of selected sites per iteration in the Pavlidis Algorithm. The x-axis shows the number of iterations and the y-axis the size of the subalignment sampled in the respective iteration. Each color corresponds to a different tree inference method used in the respective version of the algorithm.

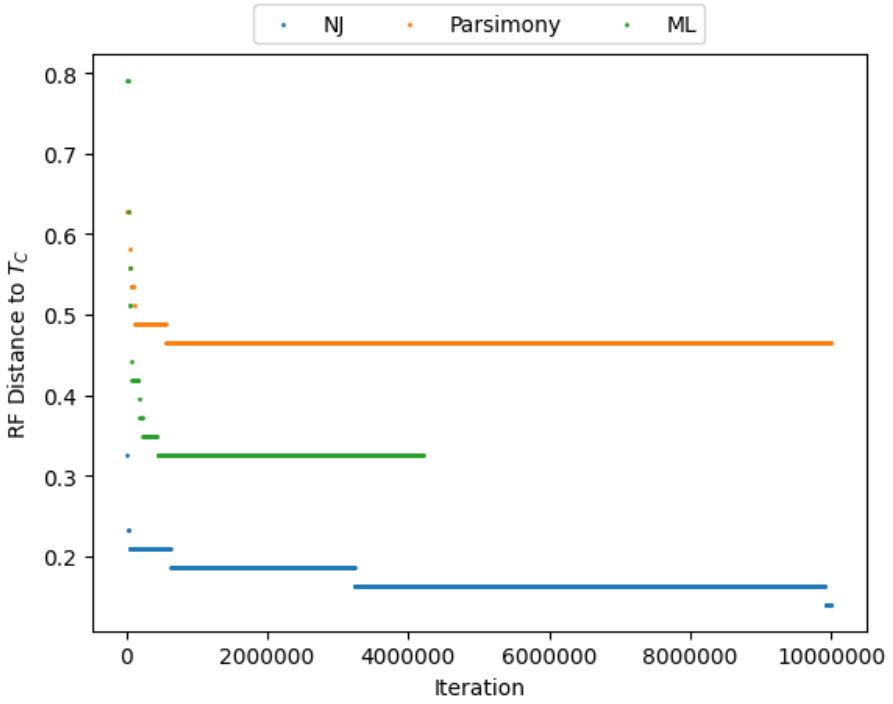


Figure 2.11.: RF distance to T_C of the inferred tree per iteration of the Pavlidis Algorithm. The x-axis shows the number of iterations, and the y-axis the RF distance to T_C. The colors correspond to the tree construction method applied in the respective version of the algorithm.

2.3.2. Result Evaluation

In the following, we conduct ML inferences on the MSAs A_{NJ}, A_{Pars} and A_{ML}. We compare the results with those obtained using the MSA $\hat{A}$. We refer to $\hat{A}$ instead of the full MSA A, since the Pavlidis algorithm only uses the informative sites.
Note, that in case of $\mathrm{Alg}_{\mathrm{ML}}$ the same tree inference method is used both in the algorithm and for evaluation, hence some of the following observations are a straight-forward consequence that simply proves the correct functionality of $\mathrm{Alg}_{\mathrm{ML}}$. Nevertheless, the analyses show relationships to results obtained from other versions of the Pavlidis Algorithm.

First, we analyze the likelihood distribution in a given set of trees. We investigate, whether trees with a low RF distance to T_C yield a comparatively high likelihood. We expect such an observation to indicate, that the respective MSA contains signal supporting T_C. In order to assess the alignment A_{NJ}, we examine the trees resulting from $\mathrm{Alg}_{\mathrm{NJ}}$ by sampling every 10 000th iteration. For each tree, we determine its likelihood regarding $\hat{A}$ and regarding A_{NJ}. To evaluate the results of $\mathrm{Alg}_{\mathrm{Pars}}$ and $\mathrm{Alg}_{\mathrm{ML}}$, we proceed in the analogous way, always using $\hat{A}$ as a reference. For all three sets of sampled trees, the likelihoods with respect to $\hat{A}$ are positively correlated with the RF distances of the respective tree to T_C. The Pearson correlation coefficient is 0.53 for $\mathrm{Alg}_{\mathrm{NJ}}$ (p-value $\ll 10\mathrm{e}-100$), 0.44 for $\mathrm{Alg}_{\mathrm{Pars}}$ (p-value $\ll 10\mathrm{e}-100$), and 0.56 for $\mathrm{Alg}_{\mathrm{ML}}$ (p-value $\ll 10\mathrm{e}-100$). Thus, trees less similar to the consensus tree are favored by ML computations on the MSA $\hat{A}$. Using A_{NJ} or A_{ML} instead, leads to a negative correlation of the obtained likelihoods and the RF distances to T_C. The Person correlation coefficients are -0.13 (p-value $\ll 10\mathrm{e}-100$) for A_{NJ} and -0.72 (p-value $\ll 10\mathrm{e}-100$) for A_{ML}. Thus, the selection of sites results in better likelihoods for trees closer to T_C. These results suggest that the algorithm successfully selects sites containing signal supporting the consensus tree. Regarding $\mathrm{Alg}_{\mathrm{Pars}}$, we do not observe such an improvement. The likelihoods regarding A_{Pars} show a slightly positively correlation to the RF distances to T_C, with a Pearson correlation coefficient 0.17 (p-value $\ll 10\mathrm{e}-100$).

To attain a deeper understanding of the quality of the resulting subalignments, we use them as an input for inferring trees with RAxML-NG. On each of $\hat{A}$, A_{NJ}, A_{Pars}, and A_{ML}, we perform 20 tree inferences using 10 random and 10 parsimony-based starting trees, yielding a tree T_{best} with the highest log-likelihood among the 20 inferred trees.

By $T_{\mathrm{best}}(A_{\mathrm{NJ}})$ we denote T_{best} resulting from tree inferences on A_{NJ}. Computing log-likelihoods with respect to A_{NJ}, we observe a major difference when comparing the scores for the sampled trees from $\mathrm{Alg}_{\mathrm{NJ}}$ to the score of $T_{\mathrm{best}}(A_{\mathrm{NJ}})$ with the latter one being substantially better. The difference between log-likelihoods is smaller for the results of $\mathrm{Alg}_{\mathrm{Pars}}$. There are trees sampled from $\mathrm{Alg}_{\mathrm{ML}}$ with a low RF distance to T_C, whose likelihood with respect to A_{ML} is almost as good as the likelihood of $T_{\mathrm{best}}(A_{\mathrm{ML}})$. As argued above, this is the consequence of using the same inference method for both, algorithm, and evaluation.

Furthermore, we examine the similarity of trees resulting from the tree searches compared to the consensus tree. $T_{\mathrm{best}}(\hat{A})$ yields a high RF distance of 0.88 to T_C. $T_{\mathrm{best}}(A_{\mathrm{NJ}})$ and $T_{\mathrm{best}}(A_{\mathrm{Pars}})$ are only slightly closer with RF distances of 0.77 and 0.81 respectively. Among the inferred trees, $T_{\mathrm{best}}(A_{\mathrm{ML}})$ stands out with an RF distance of 0.44. We assume that a tree inference on a subalignment containing mainly sites whose signal supports the consensus tree, retrieves a tree that is close to it. Under this assumption, observations related to $\mathrm{Alg}_{\mathrm{ML}}$ only indicate, that signal for T_C can be retrieved from the particular subalignment.

It is worth noting that the RF distances of the trees T_{best} are higher than those inferred during the Pavlidis algorithm, even though the respective trees are based on the very same subalignment. For neighbor joining and parsimony, this is a result of a divergent behavior of the tree inference heuristics compared to ML. However, we observe differences in distances when using ML as well. $T_{\text{best}}(A_{\text{ML}})$ is the best tree out of 20 tree searches, however, the tree in the Pavlidis Algorithm results from a single tree inference only. The fact that the RF distances of these trees to T_C differ, indicates instabilities, which we further investigate in the following Section 2.3.3.

2.3.3. Instabilities

Here we describe our experiments for assessing the stability of the results obtained from the Pavlidis algorithm. We perform a separate analysis for each of the three algorithm versions.
For Alg_{NJ}, we generate 100 bootstrapped alignments of A_{NJ}. For each bootstrap replicate, we infer a tree using NJ. Further, we compare these trees to the consensus tree T_C. We observe, that the average RF distance to T_C is 0.77 with a standard deviation of 0.07. The tree inferred on the original MSA, A_{NJ} thus turns out to be an outlier, with an RF distance of 0.14 to T_C.

As parsimony is non-deterministic, we can obtain distinct parsimony trees with exactly identical scores for the same MSA. We repeatedly apply the parsimony algorithm to A_{Pars}, resulting in a set of 500 unique trees. Considering the RF distances of these trees to T_C, we obtain an average distance of 0.71 and a standard deviation of 0.05. The tree inferred during Alg_{Pars} yields an RF distance of 0.47 to T_C, which again turns out to be an exceptionally low value.

To assess the stability of Alg_{ML}, we infer a total of 400 trees for A_{ML} using RAxML-NG. We perform 200 tree searches under BIN and BIN+G respectively. Under each model, we initiate the tree search with 100 random and 100 parsimony starting trees. Again, we compare the resulting trees to T_C, observing an average RF distance of 0.50 with a standard deviation of 0.06. Within Alg_{ML}, a tree is inferred whose RF distance to T_C is 0.33, a value residing in the tail of the distribution.

Overall, we observe substantial instabilities, regardless of which algorithm is used for tree inference. Each of the resulting MSAs A_{NJ}, A_{Pars}, and A_{ML} is an outlier in the analyzed distribution. In light of these observations, the results obtained from the Pavlidis Algorithm do not allow us to conclude that there is a clear signal in MSA A that would support T_C.

2.4. Mixture Model

In the context of maximum likelihood based methods, the phenomenon that certain sites evolve differently is often captured by means of mixture models ([57]). Under a mixture model, the likelihood of a site is determined as the weighted sum of its respective per-site likelihoods under the different models considered. The respective weights are estimated from the data by optimization. Like this, it is possible to model heterogeneity without an explicit partitioning of the sites.
In the following, we introduce a new mixture model. We investigate, whether it is suitable to infer trees from the MSA A, which are closer to T_C. For this purpose, we consider a fixed set of trees. We calculate each tree's log-likelihood score under the new model and compare the results to the distribution of the log-likelihoods under the BIN+G model. For a more independent assessment, we extend our analysis to a simulated MSA.

Our model aims to capture, that an MSA contains sites, which provide a signal for a known tree, in our case T_G. The aim is to decrease the impact of these sites on the inferred tree without explicitly having to identify them. According to this mixture model, the per-site log-likelihood $\log L_T^*(i)$ for a site i and a tree T is computed as stated in Equation (2.1). $L_T(i)$ is the per-site likelihood of site i with respect to T, $L_G(i)$ corresponds to the per-site likelihood with respect to T_G. The weight parameter w is optimized over the entire alignment.

$$\log L_T^*(i) = \log((1-w) \cdot L_T(i) + w \cdot L_G(i)) \tag{2.1}$$

2.4.1. Evaluation

In order to evaluate the effectiveness of the mixture model, we make use of a set of trees occurring as intermediate results during 400 ML searches on A (for details on the experiment see Section 2.2.5). For each tree T, we calculate its log-likelihood $\log L_T$ under BIN+G as well as the log-likelihood $\log L_T^*$ under the newly introduced model, both with respect to the MSA A. We examine, whether the new model leads to a relative improvement of the log-likelihoods for trees, which exhibit a lower RF distance to T_C.

The results are illustrated in Figure 2.12. The x-axis indicates the RF distances of the examined trees to T_C, the y-axis indicates the trees' log-likelihoods regarding A, both under the new mixture model (blue) and under the regular model (orange). For each considered tree T, it holds that $\log L_T^* \geq \log L_T$ which results from the definition of the mixture model. The rank correlation of the distributions of $\log L$ and $\log L^*$ is 1.00. Sorting the trees by their log-likelihood hence results in the same order for both models. Thus, we cannot observe any improvement in a way that trees closer to the consensus tree are being favored when we use this mixture model.

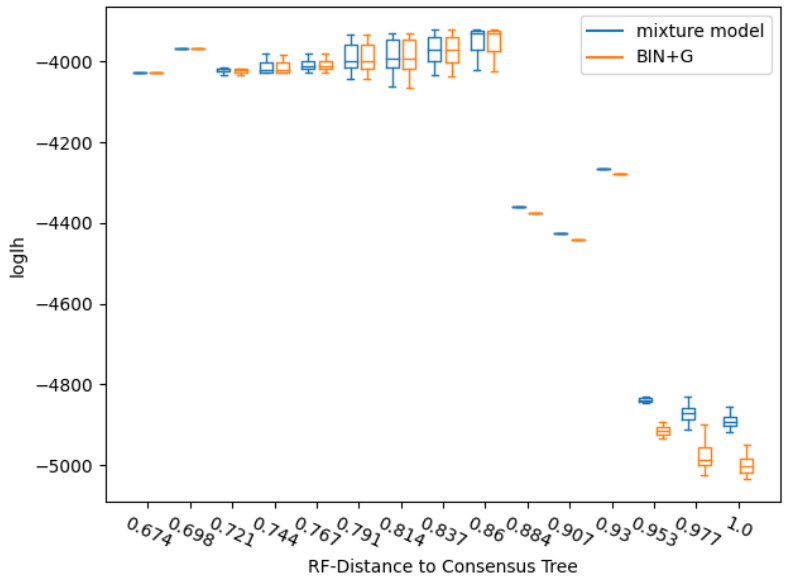


Figure 2.12.: Analysis on a set of trees

Figure 2.13.: Effect of the mixture model for MSA A. It refers to a set of trees occurring as intermediate results during 400 ML tree inferences on A. The x-axis indicates the RF distances of the examined trees to T_C, the y-axis indicates the trees' log-likelihoods regarding A. This is a boxplot, that is, the boxes represent the inter-quartile range and the horizontal bar gives the median. The results obtained under our mixture model are drawn in blue, the results obtained under BIN+G in orange.

The mixture model is based on the assumption, that horizontally evolving sites yield a higher per-site likelihood for the geographical tree, while the per-site likelihood of the consensus tree is higher for vertically evolving sites. However, the per-site likelihoods of T_C and T_G with respect to A are either correlated or extremely low, as depicted in Figure 2.3a. In light of these results, it is not surprising, that our approach is not effective.

Concerning the MSA A, we have no certainty about the extent to which sites evolve according to T_C or according to T_G. For a better assessment of the effectiveness of the mixture model, we perform additional analyses on simulated data. Using Alisim [51] we generate simulated MSAs for T_G and T_C under the GTR+F0 model, each of them comprising 2000 sites. We draw sites from these MSAs at random and mix them according to varying proportions in order to generate MSAs for our experiments. We consider MSAs whose sites are taken at 10%, 25%, 50%, and 75%, respectively, from the MSA simulated for T_G and for the remaining proportion from the MSA simulated for T_C. In the results we cannot observe the intended effect of the mixture model. However, the per-site likelihood of T_C and T_G regarding the simulated MSAs are also correlated. The exact reason for this behaviour remains unclear, but it could be related to similarities of T_C and T_G. This makes the simulated MSAs unsuitable for assessing the mixture model and hence, this additional analysis does not lead to any further insights.

Overall, for the considered trees and MSAs, our new mixture model does not appear to be suitable to handle MSAs which comprise a mixture of sites evolving according to different trees. For a final assessment of the approach, more detailed experiments are required. Since we have no evidence of the effectiveness of the mixture model, the results do not allow us to draw any conclusions concerning the MSA A.

2.5. Conclusion

In this chapter, we conducted a thorough case study on the morphosyntactic MSA A. We investigated several approaches to retrieve signal from the MSA A, which supports the consensus tree T_C. Related to this, we aimed to identify the sites in A, which evolve horizontally, that is, according to the geographical tree. We leveraged several established methods (see Section 2.2) and conducted experiments with novel ideas (see Section 2.3 and Section 2.4). Unfortunately, we were not able to retrieve the relevant signal to distinguish horizontal and vertically evolving sites in the MSA A. We conclude, that this signal is most probably simply not present in the data.
Yet, the issues we explore in this case study are also of interest in a broader context. If phylogenetic methods are applied in linguistics, where reference trees are available, this requires methods to test, to which extent an MSA contains signal for a known tree. Further, approaches are required to handle information in the MSA, which is not related to vertical evolution. Assessing whether the introduced methods are suitable to handle these issues remains subject to further analyses on additional data.

3. Data Analysis

Tools for phylogenetic inference (like RAxML-NG) are mainly developed and optimized for biological MSAs, especially for DNA and protein sequence data. We are able to construct MSAs containing linguistic data, which are in a suitable input format for such tools. However, it is an open question, to which extent these MSAs differ from biological MSAs and whether the differences impact the behaviour of heuristic tree search strategies.

In this context, we conduct analyses on a plethora linguistic MSAs and on a set of biological morphological MSAs, which we use as a reference. In Section 3.1 we introduce the data sets examined and provide information about their origin.
All MSAs under study are based on categorical data. Therefore, we can choose to represent a data set by a binary MSA or by a multi-valued MSA (see Section 1.3.5). We first investigate the impact of this representation on the tree inferences. For this purpose, we consider biological and linguistic data sets separately (see Section 3.2 and Section 3.3 respectively). In Section 3.4, we subsequently compare linguistic and biological data, restricting ourselves to binary MSAs.

3.1. Examined Data

Here we describe the data on which we perform our analyses. We indicate where we obtain the data sets and in what format they are provided to us. We first discuss biological data (see Section 3.1.1) and then linguistic data (see Section 3.1.2). Among the linguistic data sets, there are also pairs of MSAs extracted from the same primary data source. In Section 3.1.2.1 we examine these pairs in detail.

3.1.1. Biological Data

We examine 379 MSAs containing biological morphological data, provided in TreeBase [59, 77] in multi-valued representation. Additionally, there are 122 binary MSAs available. Such MSAs can either be the binary representation of a multi-state matrix or a multi-valued representation of a single-state matrix, in which not more than two different values occur for all characteristics. As we are unaware of the assembly history of these data sets, we are not able to categorize them, such that the following analyses could be conducted in a meaningful way. Hence, we decide to exclude the binary morphological MSAs.

3.1.2. Linguistic Data

Several databases containing biological MSAs are available online [6, 41, 59, 77]. However, there is no equivalent in the area of language phylogenetics, which would be suitable for our purposes. Thus, we conducted extensive research in order to collect linguistic data from different sources. Like this, we build a database similar to RAxML-Grove [41], on which we conduct our analyses. We obtain the majority of data sets in linguistic databases. For each data set, our database contains a binary MSA and, if possible, a mulit-valued MSA as well (for details see Section 3.3).
Lexibank [48] is inspired by the GenBank database [6], but instead of nucleotide sequences, it contains cross-linguistic lexical data. Cross-linguistic means, that the data sets contain information concerning multiple languages [33]. Lexibank is focused on lexical data, hence the data sets are mainly standardized word lists collected from various independent sources [48].
We obtain further data sets from supplementary material provided for the book "Sequence Comparison in Historical Linguistics" [50].
In both data sources, the data sets are standardized as specified by the Cross-Linguistic Data Format (CLDF) [33]. If all necessary information is provided, we can retrieve MSAs from these representations as described in Section 1.3.5. According to CLDF, two main types of linguistic data are distinguished. There is congnate data, provided in the form of wordlists. and structural data, containing morphophonological or morphosyntactic information.
Some additional data sets in CLDF are available independent of Lexibank and Sequence-Comparison (ewave [43], Tuled [35], sails [55] and wals [20]).
In contrast to Lexibank, the phlorest database [37] is more focused on language phyogenetics. It therefore contains the MSAs from primary sources of the respective data sets (if available).
Further, we obtain linguistic MSAs from the supplementary material of a paper by *Jäger* [42]. The authors provide cognate and sound class data for phylogenetic inference (available online: `https://osf.io/cufv7/`). The data originates from another database, the Automated Similarity Judgment Program (ASJP) [78].
DiACL [13] contains morphosyntactic data for three language families. We derive binary MSAs by manually converting of the files. Regarding the data provided by *Carling* [14], we proceed in the same way.
Our database also contains the MSAs derived from WALS [20] and the material provided by *Bouckaert et al.* [10], which we use in the case study in Chapter 2. More detailed information on the structure of the database is provided in Appendix B.

3.1.2.1. Duplicate Analysis

Lexibank and phlorest are two databases designed for different purposes. However, they overlap in several data sets (for details see Appendix B.1). This means that the respective MSAs in our database result from the same primary source. However, the MSAs we derive from CLDF data in Lexibank are not identical with those available in phlorest. First, this may be related to updates in the data sets. Second, it is also possible, that the authors of the primary sources annotate or interpret the original data differently. For example, they may group words into different cognate classes [21] or select synonyms (see Chapter 4). The MSA which is obtained by the linguists and provided in phlorest subsequently differs from the one we construct.

In the following, we examine the differences among MSAs generated from the same primary data source. For 14 out of 18 considered pairs, we are able to find a mapping for the taxon names, which means we can ensure, that both MSAs comprise the same language sets. In two cases, the Lexibank data set contains more languages, in two other cases, the MSAs contain languages which are not contained in the other one.
We further observe that there are duplicate pairs which differ substantially concerning the number of sites and/or the number of patterns, but there is no clear connection between these differences. Hence, one MSA in a pair cannot simply be a subalignment of the other one.

We use RAxML-NG to perform 100 independent tree inferences on each MSA of such duplicate pairs. For each MSA, we consider the tree with the best likelihood resulting from all tree inferences. For each duplicate pair, we determine the RF distance between the trees inferred like this on the respective MSAs. For 9 of the pairs considered, we observe an RF distance < 0.1. Hence, even though the MSAs differ, they still appear to contain analogous phylogenetic signals. For 8 duplicates, the RF distances are in a range between 0.1 and 0.3, and for one pair we even observe an RF distance of 0.67 between the inferred trees. Hence, there are also duplicate pairs in which MSAs clearly differ in the information they contain.

From our analysis, no clear conclusions can be drawn about how MSAs from the same primary sources are related. Therefore, we decide to keep both MSAs of each duplicate pair in our database. Furthermore, the analysis shows that inconsistency of linguistic data sets is a problem that should not be neglected.

3.2. Properties of Biological Binary and Multi-Valued MSAs

Let $\mathcal{D}_{\text{bio}}$ be the set of all considered biological data sets. For each data set $D \in \mathcal{D}_{\text{bio}}$ we can construct a binary MSA A and multi-valued MSA A^*. We denote D by the tuple (A, A^*). Here, we first investigate, how the properties of A and A^* are related. We further analyze the impact of the representation and of the related model on the tree inference with RAxML-NG. From these results, we aim to derive recommendations for tree inferences on categorical data.

A characteristic in a categorical data set admits a certain number of possible values. It is represented by the corresponding number of sites in the binary MSA but only by a single site in the multi-valued MSA. For a data set $D = (A, A^*)$, the number of sites in A hence increases compared to the number of sites in A^*. The increase factor corresponds to the average number of values per characteristic in the data set.
The number of different values for a characteristic also corresponds to the number of symbols occurring in the respective site in the multi-valued MSA. For the data sets in $\mathcal{D}_{\text{bio}}$, most of the characteristics admit only 2 or 3 different values. Hence, each multi-valued MSA binary for the largest proportion of its sites.
A site is said to be invariant if it is fully conserved [39], that is, it has the same value for all taxa. If there is a characteristic, such that the site representing it in A^* is invariant, all sites representing it in A are invariant as well. In the binary MSA, an invariant site additionally arises, when there is a value for a characteristic that never occurs. Hence, there are more invariant sites in the binary MSAs than in the multi-valued MSAs.
Further, we consider the number of patterns, that is the number of unique sites [39] in the MSAs. Let c_1 and c_2 be two characteristics such that the same pattern occurs in the sites representing them in A^*. These characteristics also yield identical representations in the binary MSA A, consisting of as many sites as there are possible values for the characteristics. A site corresponding to a certain value of c_1 in A is identical to the site representing the respective value of c_2. In A, two sites associated with different characteristics can also exhibit the same pattern even if the sites representing these characteristics in A^* are different. Relative to the number of sites, a binary MSA can therefore contain fewer patterns than the corresponding multi-valued MSA.

We define the *entropy* of an MSA as the average over the per-site entropies computed according to Shannon [67]. For a data set $D = (A, A^*) \in \mathcal{D}_{\text{bio}}$, the ratio of the entropy of A^* and the entropy of A is correlated with the average number of values per characteristic. The Pearson-Correlation-Coefficient of $\mathcal{D}_{\text{bio}}$ is 1.00 (p-value $< 10e - 105$). In a binary MSA, the same amount of information is spread over a larger number of sites than in the respective multi-valued MSA. It is a sparser representation, which differs more from an equal distribution, yielding a maximum entropy. Thus, it has a lower entropy than the multi-valued representation of the same information. This effect becomes stronger when the characteristics yield larger sets of possible values.

We now examine, how the tree inference behaviour with RAxML-NG differs depending on the setup. A setup M comprises a model, which is used for the tree inference, and a datatype. The datatype indicates, whether the tree inference for a data set $D = (A, A^*) \in \mathcal{D}_{\text{bio}}$ is conducted on A or A^*. We denote each setup by the name of the model used. We consider the binary MSAs under the BIN model. For the multi-valued MSAs, we examined both the GTR and MK models. GTR is more flexible, but can be prone to overparameterization, especially for MSAs with a high number of symbols (see Section 1.4.3.1). For every setup M and every data set $D \in \mathcal{D}_{\text{bio}}$, we perform 100 tree inferences. We denote the best-known ML tree by $T_M^{\text{best}}(D)$.

Based on the tree inferences, we determine the difficulty score (see Section 1.4.3.4) for each MSA under each of the three distinct setups. The respective distributions are depicted in Figure 3.1. Under GTR, the scores are higher (that is, inferences are more difficult) than under MK and under BIN. Thus, the difficulty of a data set is not only related with the information contained, but also with the representation and the model which is used for the tree inference. On the one hand, the high difficulty scores under the GTR model may be a consequence of overparameterization. On the other hand, the BIN model may not suffice to capture all relationships in the data set. In the following analysis we determine the setup under which the resulting difficulty score most accurately describes the data itself.

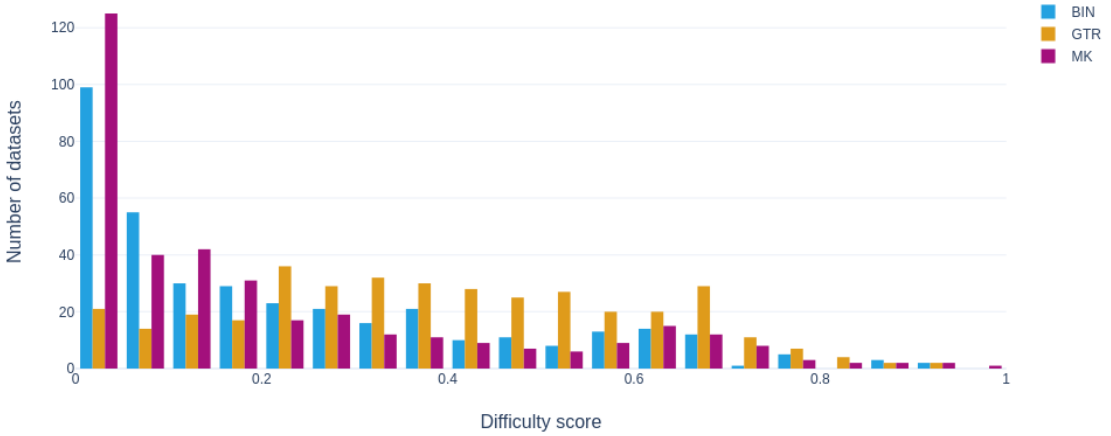


Figure 3.1.: Distribution of difficulty scores under different setups
The x-axis indicates the difficulty score, the y-axis the number of data sets with the respective score. The bar's colors correspond to the setup under which the evaluation takes place.

For each data set $D \in \mathcal{D}_{\text{bio}}$ we compute the pairwise RF distances between $T^{\text{best}}_{\text{BIN}}(D)$, $T^{\text{best}}_{\text{GTR}}(D)$, and $T^{\text{best}}_{\text{MK}}(D)$. Comparing the trees inferred under BIN and MK, the average RF distance over all data sets is 0.20. It is 0.38 for BIN and GTR and 0.40 for MK and GTR. This shows, that the setup affects the tree inferences. In particular, GTR yields trees, which differ substantially from those inferred under the two other setups.

For each pair of setups $M_i, M_j \in \{\text{BIN}, \text{GTR}, \text{MK}\}, M_i \neq M_j$, we consider the best tree found under M_j, evaluate its log-likelihood in M_i and compare it to the log-likelihood of the best tree found under M_i itself. For this comparison, we introduce a metric we call the cross difference. Let $\log \text{L}_M(T)$ be the log-likelihood of a tree T under setup M. For a data set $D \in \mathcal{D}$ and two setups M_i, M_j, $M_i \neq M_j$, we define the cross difference as follows:

$$\text{diff}_{M_i,M_j}(D) := \frac{\log \text{L}_{M_i}(T^{\text{best}}_{M_i}(D)) - \log \text{L}_{M_i}(T^{\text{best}}_{M_j}(D))}{\log \text{L}_{M_i}(T^{\text{best}}_{M_i}(D))} \tag{3.1}$$

We provide relative values for the cross differences, since the data sets yield a broad range of absolute log-likelihood values ($\log \text{L}_M(T^{\text{best}}_M(D))$ ranges between -1.34 and -27840.68). As we report the log-likelihoods instead of the likelihoods themselves, the observed effects are greater than the relative differences might suggest.

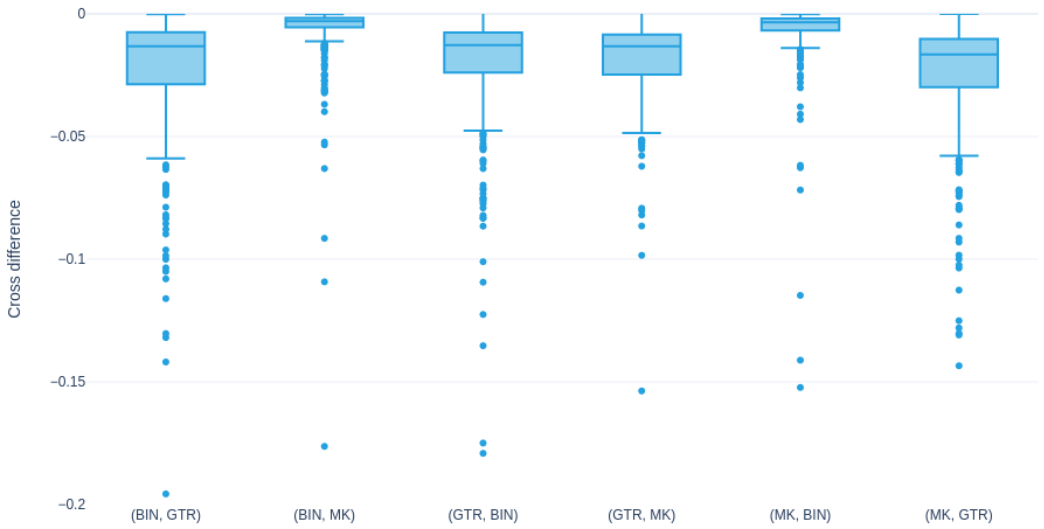


Figure 3.2.: Distribution of cross differences in $\mathcal{D}_{\text{bio}}$
The y-axis indicates cross differences. Each boxplot illustrates the distribution of $\text{diff}_{M_i,M_j}(D)$ for the setups (M_i, M_j) indicated on the x-axis. The boxes represent the interquartile range, and the horizontal bar indicates the median.

Figure 3.2 illustrates, how the cross differences are distributed for the data sets in $\mathcal{D}_{\text{bio}}$. Each box plot refers to $\text{diff}_{M_i,M_j}(D)$ for the setups M_i, M_j indicated on the x-axis. The y-axis indicates the cross differences. We observe that the resulting differences are always negative. Hence, for a fixed setup $M_i \in \{\text{BIN}, \text{GTR}, \text{MK}\}$, any tree inferred under another setup M_j is never better than the tree inferred in M_i itself. Conversely, however, we observe, that there are data sets for which the tree resulting from the inference with M_j admits a substantially worse log-likelihood when evaluated under the model of M_i. The effect is strongest, when $\text{GTR} \in \{M_i, M_j\}$.

Although the observations regarding RF distances and cross differences have a similar structure, the metrics are not correlated at all. Hence, we conclude, that the different setups induce substantially different likelihood distributions in tree space. This justifies why the difficulty scores differ depending on the setup.

We also consider the Akaike Information Criterion (AIC) score [2] suitable for the comparison of different models. In phylogenetics, the criterion provides an estimate of the information, which is lost when choosing a specific model to represent the evolutionary process. To minimize this loss, a model must neither be too simple nor too complex. A lower AIC score indicates, that the result obtained with the respective model is superior. We calculate an AIC Score for the result of each tree inference. For each setup M and every set D we compute $\mathrm{AIC}_M(D)$, the average AIC score over the 100 tree inferences performed on D under M. We observe $\mathrm{AIC}_{\mathrm{GTR}}(D) < \mathrm{AIC}_{\mathrm{MK}}(D) < \mathrm{AIC}_{\mathrm{BIN}}(D)$ for 375 out of 379 data sets in $\mathcal{D}_{\mathrm{bio}}$. For 3 data sets we observe $\mathrm{AIC}_{\mathrm{GTR}}(D) > \mathrm{AIC}_{\mathrm{MK}}(D)$ and for one data set $\mathrm{AIC}_{\mathrm{MK}}(D) > \mathrm{AIC}_{\mathrm{BIN}}(D)$. It follows, that using multi-valued MSAs and the GTR-Model is the setup, which yields the best model fit.

Relating the AIC scores to our observation of difficulty scores depending on the setup, we conclude, that the high difficulty scores observed with GTR do not result from over-parametrization but merely properly capture the data set properties. We further derive the recommendation to transform categorical data into multi-valued representation (if possible) and to perform tree inferences under the GTR model. If GTR is technically not feasible due to high run times, using MK instead is preferable to performing a tree inference on the binary representation of the data set.

3.3. Properties of Linguistic Binary and Multi-Valued MSAs

In this section, we investigate the impact of the representation of MSAs for linguistic data. In contrast to biological data, linguistic data sets usually have a multi-state matrix in their original representation, meaning that more than one value can be assigned to a taxon / characteristic pair. Constructing a multi-valued MSA (that is an MSA containing more than two different symbols) is therefore only possible under certain restrictions. For this reason, we limit ourselves to a brief analysis confirming the results obtained for biological data.

For the construction of multi-valued MSAs, we use the approach introduced in Section 1.3.5 with thresholds g and h. Based on preliminary experiments we decide to fix $g = 0.1$. In order to obtain a multi-valued MSA for a set of data sets that are sufficient for a meaningful analysis, we determine a value for h by experimental exploration.
For some data sets, we are only given the binary MSA A, but not the original matrix M or any other piece of information, from which we could retrieve, which sites in the binary MSA belong to the same characteristic. Hence, it is not possible to find a multi-valued MSA for these data sets. Those, for which all necessary information is available, are either supplied in CLDF (mainly in Lexibank) or retrieved from ASJP.
Several CLDF data sets exhibit a low proportion of multi-state characteristics, that are characteristics for which there is at least one multi-state entry in the categorical data matrix. With $h \coloneqq 0$ we are able to find a multi-valued representation for 31 of 86 available data sets. However, 9 of them have at most two different states. Thus, they are in fact binary and we do not investigate them further.
In case of the data sets retrieved from ASJP, the degree of multi-state characteristics is higher. We therefore set $h \coloneqq 0.1$ and obtain multi-valued MSAs for 18 out of 65 data sets. Note, that for these data sets, the degree of information loss is higher.

Among data sets for which we are able to retrieve a multi-valued MSA, the maximum number and the average number of values per characteristic tend to be higher than in biological data sets. This leads to a higher number of distinct symbols in the MSAs. The multi-valued format used in RAxML-NG can represent MSAs with up to 64 different symbols, but the inferences become highly time-consuming with an increased number of symbols. There are 8 MSAs with ≥ 15 symbols, which we therefore exclude from the following analyses. For the remaining 35 MSAs, we repeat the same experiments as for $\mathcal{D}_{\text{bio}}$.

We make similar observations for linguistic data as for biological data in the previous section. However, some trends are less clear. This may be due to the loss of information in the multi-valued representation. In addition, the number of examined data sets is too small to draw firm conclusions.

Considering the AIC Scores, we again observe that $\text{AIC}_{\text{GTR}} < \text{AIC}_{\text{MK}} < \text{AIC}_{\text{BIN}}$ for all but one data set. This implies, that, if we can construct a multi-valued MSA, using GTR with this MSA is the best fitting model. It is hence worth to consider in future work, whether it is possible to represent categorical data with multi-state matrices in a multi-valued manner. Probabilistic MSAs (see Section 4.2) are a possible approach to deal with this issue.

3.4. Comparison of Biological and Linguistic Data

In the following, we compare biological and linguistic data. We only use the MSAs in their binary representations, because for most of the linguistic data sets we are not able to construct a multi-valued MSA. We examine basic data properties and further investigate differences in the context of tree inferences with RAxML-NG.

First, we compare the MSAs regarding their size. The majority of both biological and linguistic MSAs has 40 taxa or less. However, there are groups of larger linguistic MSAs each with a similar number of taxa among them. This is presumably related with the number of languages in language families, which are frequently examined.
The median for the number of sites in biological MSAs is 229, for the linguistic MSAs it is 634. In this dimension, the linguistic MSAs hence tend to be larger. Additionally, in the examined MSA collection, there are 4 exceptionally large linguistic MSAs with > 18000 sites.
Further, we consider the ratio of the number of patterns and the number of sites, which corresponds to the ratio of unique sites. We observe, that the ratio tends to be higher in biological MSAs. In other wordss, in linguistic MSAs, it occurs more frequently, that sites have the same pattern.

Figure 3.3 depicts, how the entropy of the considered MSAs is distributed. We observe lower values for the linguistic MSAs. As argued above, this is related with a higher average number of symbols per characteristic. We make corresponding observations when comparing these numbers for biological and linguistic data sets. Breaking this down to the different types of language data, we observe an average entropy of 0.39 for cognate MSAs, 0.43 for morphological MSAs, and 0.56 for sound class MSAs. Hence, there is a relationship between the type of linguistic information encoded and the entropy of the respective MSA. Based on the construction of sound class MSAs we expected their entropy to be lower. Further investigating this difference remains subject of future work.

For each binary data set, we perform 100 tree inferences under the BIN model using RAxML-NG. We examine the difficulty scores (see Section 1.4.3.4) determined based on these inferences, and also analyze the branch lengths of the resulting trees.
Figure 3.4 depicts how the difficulty is distributed among the considered MSAs. For linguistic and biological data, a large proportion of the MSAs yields a difficulty ≤ 0.1. With increasing difficulty, we observe a decrease in the number of biological MSAs. In case of linguistic data, however, there is a considerable proportion of MSAs with a difficulty exceeding 0.5.
If we separately consider the different types of language data, we can observe different distributions of the difficulty among the corresponding MSAs. Morphological MSAs are classified as the most difficult, with an average score of 0.58. For cognate MSAs, the average difficulty is 0.32, for sound class MSAs it is 0.19.

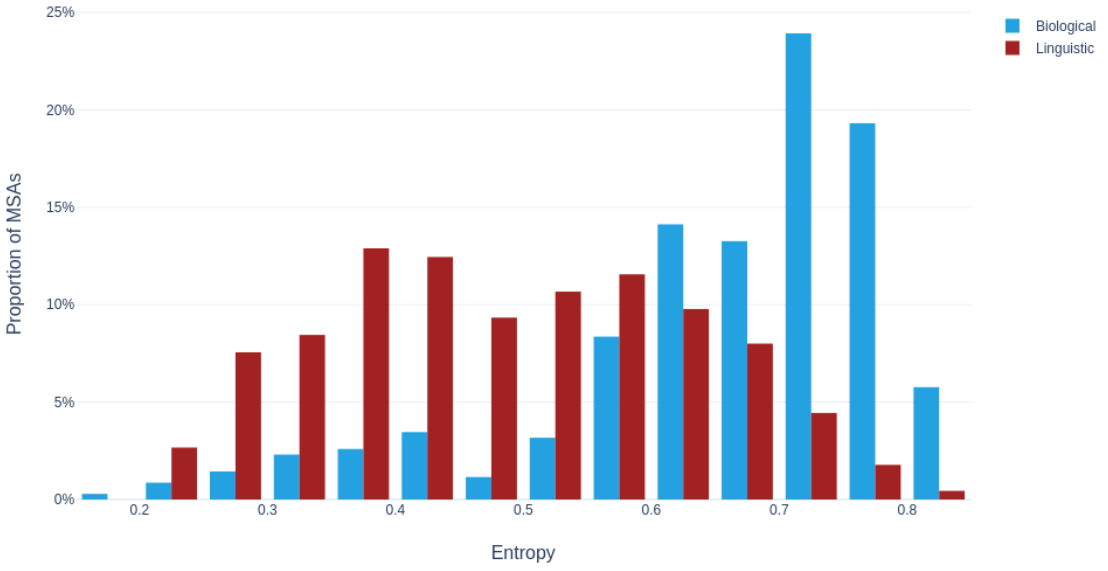


Figure 3.3.: Entropy distribution for linguistic and biological MSAs.
The x-axis indicates the entropy, the y-axis the respective proportion of MSAs. Blue bars correspond to biological, red bars to linguistic data.

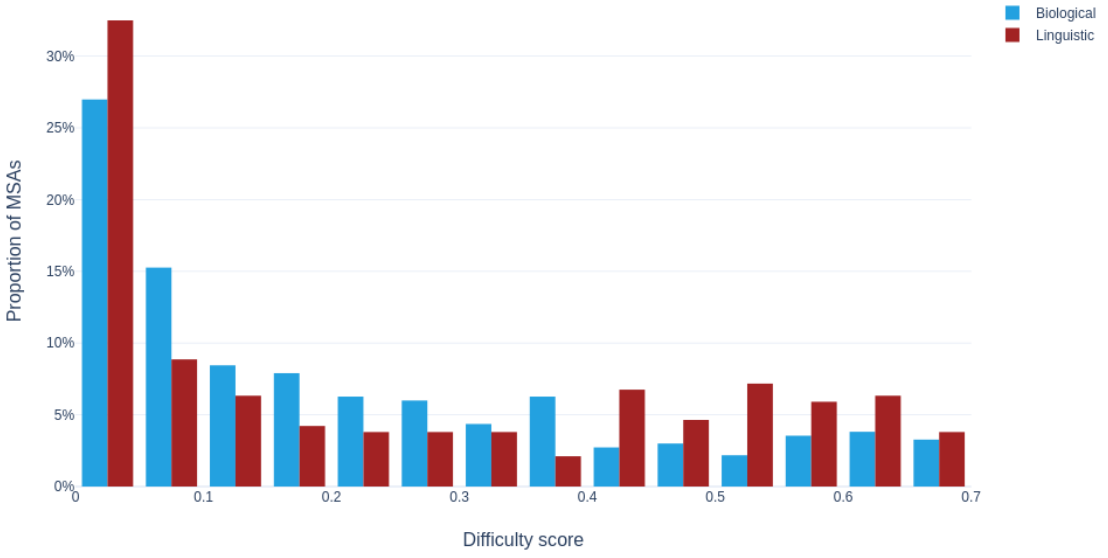


Figure 3.4.: Difficulty score distribution for linguistic and biological MSAs.
The x-axis indicates the difficulty score, the y-axis the proportion of MSAs with the respective score. Blue bars correspond to biological, red bars to linguistic data.

Additionally, we analyze how branch lengths are distributed in the trees resulting from the tree inferences. First, we consider the external branches, which connect a leaf to the rest of the tree. Figure 3.5 depicts, how the median of the branch lengths of these external branches is distributed over the best trees resulting from the tree inferences on the considered MSAs. We observe that the median is close to 0 for most of the trees inferred on biological data, while for linguistic data, the median is more evenly distributed (see Figure 3.5). This observation is presumably more related to data quality than to an evolutionary phenomenon. If an MSA has a poor resolution, there are sets of taxa, for which we are not able to infer any evolutionary relationship (e.g., because all sites are identical for these taxa). During the inference, these taxa are however still arranged in a binary subtree. The external branches in this subtree have a length close to 0, due to the lack of signal in the MSA.
For the internal branch lengths, we do not observe substantial differences between trees based on language and biological MSAs.

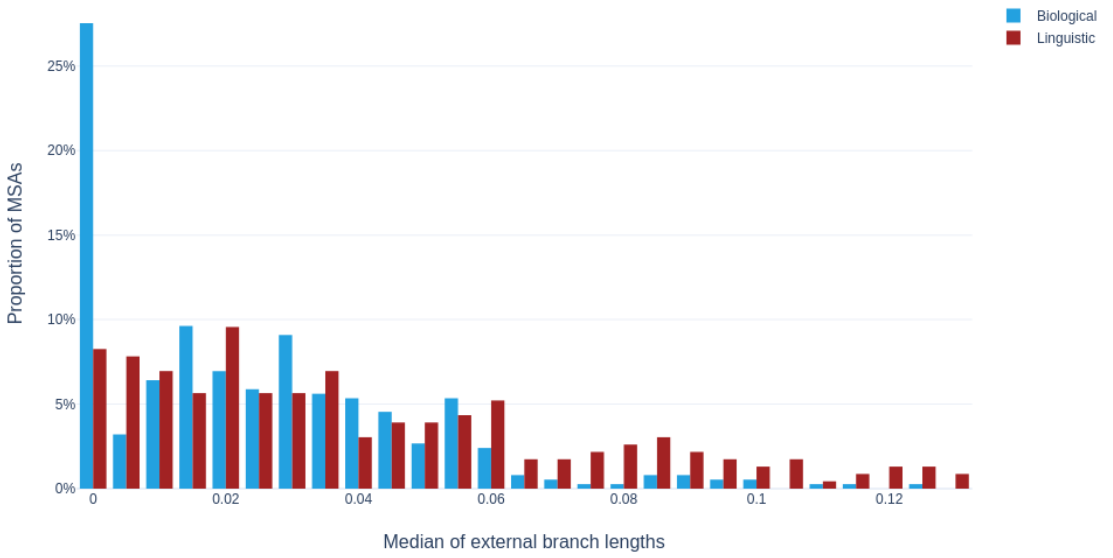


Figure 3.5.: Distribution of the median of external branch lengths for linguistic and biological MSAs
The x-axis indicates the median, the y-axis the proportion of MSAs such that the respective median occurs among lengths of the external branches in the trees inferred on base of that MSA. Blue bars correspond to biological, red bars to linguistic data.

3.5. Conclusion

In this section, we presented results of analyses we performed on different types of linguistic data, as well as on morphological data from biology. We first investigated the effects of representing a data set as a binary MSA or as a multi-valued MSA. We observed, that depending on the setup (i.e., the representation together with the model) a different distribution of likelihoods in tree space results. This leads to different difficulty scores for a data set depending on the setup. Using the AIC scores, we determined that inferring trees on the multi-valued MSA with GTR as model leads to the best model fit. However, there are categorical data sets for which we are only able to obtain a multi-valued MSA with a loss of information. Developing an alternative representation for these data sets remains subject of future work.

We further investigated how linguistic and biological data differ. We observed that, except for some more difficult biological data sets, the difficulty scores are similarly distributed in both groups of data. A closer look at the branch lengths of the inferred trees indicates potential problems regarding data quality in biological data. This issue does not occur in linguistic data.

In terms of the entropy and the difficulty score, we observed differences between the types of linguistic data studied. MSAs containing sound class data tend to admit a higher entropy and a lower difficulty score than those containing cognate data. The average entropy of morphological MSAs lies inbetween the corresponding values of the other two data types. However, morphological MSAs are on average the most difficult.

4. Modeling Subjectivity

When working with cognate data, linguists consider several semantic concepts and determine the words describing these concepts in the languages under study (see Section 1.3.4.1). Words from multiple cognate classes can denote the same concept in a language, and linguists might select only some of these so-called synonyms when generating MSAs. The selection of synonyms is often subjective rather than based on statistical or other objective criteria such as analyzing how frequently the respective words are being used [49]. In this chapter, we investigate the influence of synonym selection on inferring phylogenies (Section 4.1) and examine a novel approach based on probabilistic MSAs (Section 4.2). We perform all analyses using an example data set published by *Dunn* [22]. Verifying our findings on a broader set of MSAs remains the subject of future work.

4.1. Impact of Selecting Synonyms on Tree Inferences

In this section, we analyze to which extent selecting synonyms impacts the trees inferred with RAxML-NG. We generate 1000 distinct MSAs by selecting 1000 distinct sets of synonyms. For each generated MSA, we infer a single phylogenetic tree using RAxML-NG, resulting in a set of 1000 trees. To quantify the impact of the selected synonyms on the tree inference, we compute all pairwise RF distances among these trees. For each tree, we consider the average of the pairwise RF distances to each other tree. Section 4.1 shows, how these average distances are distributed. We observe a peak at 0.0, indicating that there is a large subset of trees, which are very similar or even identical to each other. However, we also observe average RF distances around 0.2. A tree hence differs on average 20% from the trees inferred on MSAs based on other synonym selections. We conclude that the tree inference is sensitive to the synonym selection and that this selection should not be based on subjective decisions.

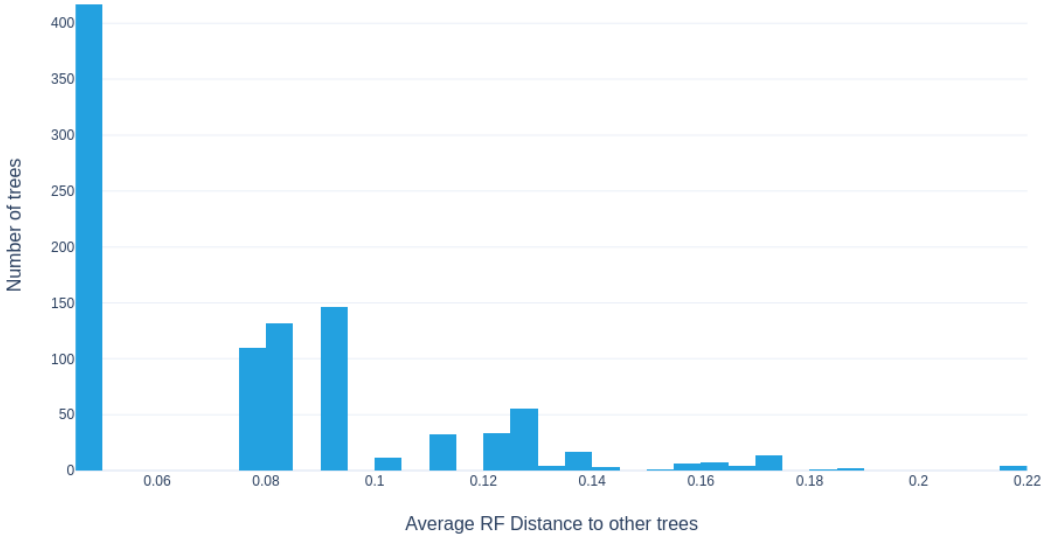


Figure 4.1.: Trees inferred on 1000 MSAs that were created by randomly selecting distinct sets of synonyms. The x-axis indicates average RF distances and the y-axis indicates the number of trees, which admit the respective average RF distance to the other trees.

4.2. Probabilistic MSAs

In the previous section, we showed for one example data set that the manual selection of synonyms results in different tree topologies inferred with RAxML-NG on the respective MSAs. Hence, we aim to find an approach to handle synonyms in cognate data, that does neither require an explicit synonym selection nor ?. For this purpose, we propose probabilistic MSAs.

4.2.1. Definition

The MSAs considered so far are all deterministic, as the underlying assumption is that a fixed symbol is observed at each site and for each taxon. In a probabilistic MSA, we instead assume that the various symbols can occur with certain probabilities. A probabilistic binary MSA A_{prob} is a matrix providing the probabilities with which we observe the symbols in Σ for each taxon and each site. In A_{prob}, missing data is represented by setting the probabilities for all symbols to 1.0 [44]. This encoding does not contain any information, and hence, the missing entries do not influence the likelihood score.

We can represent A_{prob} in a file using the CATG-Format supported by RAxML-NG [44]. The tree inference based on the probabilistic MSA differs from a standard inference in the conditional likelihood vectors at the leaves. Usually, such a vector contains a single 1.0 entry for the observed discrete value and the remaining entries are all set to 0.0. Performing an inference for A_{prob}, the conditional likelihood vectors are determined based on the probabilities provided [44].

4.2.2. Application for Synonyms

We can use probabilistic MSAs to circumvent synonym selection. For this purpose, we consider cognate data in a probabilistic context. If k synonyms exist for a concept in a language, we assume, that each of them occurs with probability $\frac{1}{k}$. Based on this consideration, we construct a binary probabilistic MSA. At a site corresponding to one of the 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.

4.2.3. Evaluation

We evaluate the introduced approach on the data set published by *Dunn* [22]. We further use T_g, a manually constructed reference tree, which we obtain by pruning a published tree by *Hammarström et al.* [40]. T_g is multifurcating, hence we evaluate the generalized quartet (GQ) distance [60] to measure dissimilarities. This topological distance metric is based on all quartets of taxa in a tree. For each quartet, we determine the topology of the subtree it induces. Comparing two trees, we obtain the GQ distance as the proportion of quartets inducing the different topologies in the trees. The advantage of the metric is that it evaluates to 0 if there are no contradictions between the inferred tree and the reference tree, even if this reference tree contains polytomies. Note that this metric is not directly comparable to the RF distance, as the values are distributed differently [70]. If two trees admit a GQ distance < 0.05, we interpret them as being highly similar. A GQ distance > 0.1 indicates substantial differences in the respective trees.

To compare the presented probabilistic approach to synonym selection, we again consider the trees corresponding to different selections of synonyms as introduced in the previous section. These trees admit an average GQ distance of 0.032 to the reference tree T_g. The maximum observed GQ Distance is 0.065.

We construct a probabilistic MSA for the given data set as described above, and perform 20 independent tree inferences using RAxML-NG. The resulting tree with the best known log-likelihood admits a GQ distance of only 0.019 to the reference tree T_g. Hence, using a probabilistic MSA does not only allow for handling synonyms without explicit selection, but also yields a tree that is topologically more similar to the reference tree.

4.3. Conclusion

Within this chapter, we used an exemplary data set to show, that synonym selection synonyms substantially impacts the trees inferred with RAxML-NG. We tested probabilistic MSAs as an approach to handle synonyms without explicitly selecting them. This circumvents biasing the phylogenetic inference by subjective decisions related to the synonym selection process. For the examined data set, we further showed, that the tree inferred on the probabilistic MSA is more similar to a manually constructed reference tree than trees based on synonym selections. Verifying our findings and further assessing the advantages of the probabilistic approach on a broader set of input data remains subject of future work. In addition, we aim to investigate alternative probability distributions for synonym occurrence. Instead of using the uniform distribution, we may, for example, derive probabilities from lexicostatistical analyses.

5. Conclusion and Future Work

5.1. Discussion

Within this thesis, we considered several aspects of linguistic data in language phylogenetics. In Chapter 2 we conducted a case study on a morphosyntactic MSA concerned with the Indo-European language family. We assumed, that this MSA contains signal supporting the consensus tree for this language family. We investigated various methods in order to reveal this potential signal. Among others, we aimed to identify the sites in the MSA, which evolve horizontally, that is according to the geographical tree. However, none of the examined approaches lead to positive results, indicating, that there is likely no signal supporting the consensus tree contained in the MSA under study. We conclude, that the respective signal is not present in the data.
In Chapter 3, we presented results from analyses of a large number of data sets containing different types of linguistic data, as well as on morphological data sets from biology. In a first analysis, we examined, to which extent the choice of MSA representation (binary versus multi-valued) and evolutionary model influences the results of the tree inference. We observed a noticeable impact on the inferred trees for both factors. We recommend to use multi-valued MSAs and the GTR model, as this leads to the best results with respect to difficulty and model fit according to our analyses. A comparison of biological and linguistic data revealed, that morphological MSAs in biology tend to exhibit poor signal for resolving the leaves in the phylogenies, which does not appear to be an issue for language data.
In Chapter 4, we investigated synonyms in cognate data. With experiments on an exemplary data set, we showed, that the selection of synonyms clearly impacts the inferred tree topologies. We concluded, that the selection process should not be based upon subjective decisions, and we proposed using probabilistic MSAs instead. We showed the potential effectiveness of this approach on the exemplary data set.

5.2. Outlook

Using methods for phylogenetic inference in the field of linguistics leads to several open questions and new requirements. The results presented in this thesis give rise to various topics for future work. The case study in Chapter 2 is an example, how to conduct phylogenetic inference together with available reference trees. Our experiments suggest, that in this context, linguists require methods for testing to which extent an MSA contains signal for a known language tree and for handling information in the MSA, which is not related to vertical evolution. In future work, we aim to develop such methods. This involves examining the effectiveness of the presented approaches on additional data sets.
For the experiments in Chapter 3 we first collected linguistic data sets in a database. We intend to further improve this database and make it publicly available. Our analyses showed, that it is beneficial for the results of the tree inferences to represent categorical data as multi-valued MSA. As discussed before, this is however currently not possible for some data sets. How these data sets could be represented instead is another question to be investigated.
In Chapter 4, we presented promising results regarding the application of probabilistic MSAs for handling synonyms. However, our observations and results are only based on a single data set. In our future research, we aim to prove the effectiveness of the approach with the help of experiments on additional data.

Bibliography

[1] R. Adams, Z. Cain, R. Assis, and M. DeGiorgio. Robust phylogenetic regression. *bioRxiv*, 2022. doi: 10.1101/2022.08.26.505424.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Selected papers of hirotugu akaike*, pages 199–213, 1998.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. The structure and function of dna. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[4] Q. D. Atkinson and R. D. Gray. Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics. *Systematic Biology*, 54(4):513–526, 08 2005. ISSN 1063-5157. doi: 10.1080/10635150590950317.

[5] S. Behjati and P. S. Tarpey. What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, 98(6):236–238, 2013.

[6] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Research*, 41(D1):D36–D42, 11 2012. ISSN 0305-1048. doi: 10.1093/nar/gks1195.

[7] S. Berger and A. Stamatakis. Accuracy of morphology-based phylogenetic fossil placement under maximum likelihood. *2010 ACS/IEEE International Conference on Computer Systems and Applications, AICCSA 2010*, pages 1 – 9, 06 2010. doi: 10.1109/AICCSA.2010.5586939.

[8] B. Bettisworth and A. Stamatakis. Root digger: a root placement program for phylogenetic trees. *BMC Bioinformatics*, 22, 05 2021. doi: 10.1186/s12859-021-03956-5.

[9] R. Borges, J. P. Machado, C. Gomes, A. P. Rocha, and A. Antunes. Measuring phylogenetic signal between categorical traits and phylogenies. *Bioinformatics*, 35(11): 1862–1869, 10 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty800.

[10] R. Bouckaert, P. Lemey, M. Dunn, S. Greenhill, A. Alekseyenko, A. Drummond, R. Gray, M. Suchard, and Q. Atkinson. Report—mapping the origins and expansion of the indo-european language family. *Science (New York, N.Y.)*, 337:957–60, 08 2012. doi: 10.1126/science.1219669.

[11] R. P. Brent. An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal*, 14(4):422–425, 01 1971. ISSN 0010-4620. doi: 10.1093/comjnl/14.4.422.

[12] C. H. Brown, E. W. Holman, S. Wichmann, and V. Velupillai. Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals*, 61(4):285–308, 2008. doi: doi:10.1524/stuf.2008.0026.

[13] G. Carling, editor. *Diachronic Atlas of Comparative Linguistics Online*. Lund University, Lund, 2017. URL `https://diacl.ht.lu.se/6`.

[14] G. Carling, editor. *Volume 1 The Mouton Atlas of Languages and Cultures, 1.* De Gruyter Mouton, Berlin, Boston, 2019. ISBN 9783110367416. doi: 10.1515/9783110367416.

[15] B. Chor and T. Tuller. Maximum likelihood of evolutionary trees is hard. In *Research in Computational Molecular Biology: 9th Annual International Conference, RECOMB 2005, Cambridge, MA, USA, May 14-18, 2005. Proceedings 9*, pages 296–310. Springer, 2005.

[16] J. Cohen. A phylogenetic analysis of morphological and molecular characters of boraginaceae: Evolutionary relationships, taxonomy, and patterns of character evolution. *Cladistics*, 30, 07 2013. doi: 10.1111/cla.12036.

[17] C. Darwin. *On the origin of species by means of natural selection.* John Murray, London, 1859.

[18] C. Darwin. *The Descent of Man, and Selection in Relation to Sex.* John Murray, London, 1871.

[19] M. J. de Smith, M. F. Goodchild, and P. A. Longley. Geospatial analysis: comprehensive guide to principles, techniques and software tools, 2018-21.

[20] M. S. Dryer and M. Haspelmath, editors. *WALS Online (v2020.3).* Zenodo, 2013. doi: 10.5281/zenodo.7385533.

[21] M. Dunn. Language phylogenies, 08 2013.

[22] M. Dunn. Cldf dataset derived from dunn's "ielex" from 2012, jul 2021.

[23] M. Dunn, A. Terrill, G. Reesink, R. Foley, and S. Levinson. Structural phylogenetics and the reconstruction of ancient language history. *Science (New York, N.Y.)*, 309: 2072–5, 10 2005. doi: 10.1126/science.1114615.

[24] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge university press, 1998.

[25] B. S. Everitt and A. Skrondal. The cambridge dictionary of statistics. 2010.

[26] J. S. Farris. Methods for computing wagner trees. *Systematic Zoology*, 19(1):83–92, 1970.

[27] J. Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, 22(3):240–249, 1973.

[28] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27:401–410, 1978.

[29] J. Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 2005.

[30] M. Fischer, L. van Iersel, S. Kelk, and C. Scornavacca. On computing the maximum parsimony score of a phylogenetic network. *SIAM Journal on Discrete Mathematics*, 29, 02 2013. doi: 10.1137/140959948.

[31] W. M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.

[32] R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, New York, NY, USA, second edition, 1987.

[33] R. Forkel, J.-M. List, S. Greenhill, C. Rzymski, S. Bank, M. Cysouw, H. Hammarström, M. Haspelmath, G. Kaiping, and R. Gray. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5:180205, 10 2018. doi: 10.1038/sdata.2018.205.

[34] S. A. Fritz and A. Purvis. Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, 24, 2010.

[35] F. F. Gerardi, S. Reichert, C. Aragon, T. Wientzek, J.-M. List, and R. Forkel. Tuled. tupían lexical database, 2022.

[36] R. Gray and Q. Atkinson. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426:435–9, 12 2003. doi: 10.1038/nature02029.

[37] S. Greenhill. phlorest. https://github.com/phlorest, 2023.

[38] S. Greenhill, C.-H. Wu, X. Hua, M. Dunn, S. Levinson, and R. Gray. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, 114, 10 2017. doi: 10.1073/pnas.1700388114.

[39] J. Haag, D. Höhler, B. Bettisworth, and A. Stamatakis. From easy to hopeless - predicting the difficulty of phylogenetic analyses. *bioRxiv*, 2022. doi: 10.1101/2022.06.20.496790.

[40] H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank. Glottolog 4.7, 2022. URL `http://glottolog.org`.

[41] D. Höhler, W. Pfeiffer, V. Ioannidis, H. Stockinger, and A. Stamatakis. RAxML Grove: an empirical phylogenetic tree database. *Bioinformatics*, 38(6):1741–1742, 12 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab863.

[42] G. Jäger. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5, 10 2018. doi: 10.1038/sdata.2018.189.

[43] B. Kortmann, K. Lunkenheimer, and K. Ehret. ewave, 2020. URL `https://ewave-atlas.org/`.

[44] O. Kozlov. *Models, Optimizations, and Tools for Large-Scale Phylogenetic Inference, Han- dling Sequence Uncertainty, and Taxonomic Validation.* PhD thesis, Karlsruher Institut für Technologie (KIT)y, 2018.

[45] M. Ladoukakis, D. Michelioudakis, and E. Anagnostopoulou. Toward an evolutionary framework for language variation and change. *BioEssays*, 44:2100216, 01 2022. doi: 10.1002/bies.202100216.

[46] M. S. Lee and A. Palci. Morphological phylogenetics in the genomic age. *Current Biology*, 25(19):R922–R929, 2015. ISSN 0960-9822. doi: 10.1016/j.cub.2015.07.009.

[47] P. O. Lewis. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Systematic Biology*, 50(6):913–925, 11 2001. ISSN 1063-5157. doi: 10.1080/106351501753462876.

[48] J.-M. List, R. Forkel, S. Greenhill, C. Rzymski, J. Englisch, and R. Gray. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9:316, 06 2022. doi: 10.1038/s41597-022-01432-0.

[49] M. List. Tossing coins: linguistic phylogenies and extensive synonymy, 02 2018. URL `https://phylonetworks.blogspot.com/2018/02/tossing-coins-linguistic-phylogenies.html`.

[50] M. List. *Sequence Comparison in Historical Linguistics*. düsseldorf university press, Berlin, Boston, 2021. ISBN 9783110720082. doi: doi:10.1515/9783110720082.

[51] N. Ly-Trong, S. Naser-Khdour, R. Lanfear, and B. Minh. Alisim: A fast and versatile phylogenetic sequence simulator for the genomic era. *Molecular biology and evolution*, 39, 05 2022. doi: 10.1093/molbev/msac092.

[52] P. H. Matthews. *The Concise Oxford Dictionary of Linguistics*. Oxford University Press, 2007. ISBN 9780191727160. doi: 10.1093/acref/9780199202720.001.0001.

[53] D. Michelioudakis, M. Ladoukakis, P. Pavlidis, A. M. Ramadanidis, M.-M. Makri, and E. Anagnostopoulou. Exploring the morphosyntactic characters of wals for the phylogenetic reconstruction of languages, 2021.

[54] B. Morel, P. Barbera, L. Czech, B. Bettisworth, L. Hübner, S. Lutteropp, D. Serdari, E.-G. Kostaki, I. Mamais, A. M. Kozlov, P. Pavlidis, D. Paraskevis, and A. Stamatakis. Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. *Molecular Biology and Evolution*, 38(5):1777–1791, 12 2020. ISSN 1537-1719. doi: 10.1093/molbev/msaa314.

[55] P. Muysken, H. Hammarström, O. Krasnoukhova, N. Müller, J. Birchall, S. van de Kerke, L. O'Connor, S. Danielsen, R. van Gijn, and G. Saad, editors. *South American Indigenous Language Structures (SAILS)*. Max Planck Institute for the Science of Human History, Jena, 2016. URL `https://sails.clld.org`.

[56] L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1):268–274, 2015.

[57] M. Pagel and A. Meade. A Phylogenetic Mixture Model for Detecting Pattern-Heterogeneity in Gene Sequence or Character-State Data. *Systematic Biology*, 53(4): 571–581, 08 2004. ISSN 1063-5157. doi: 10.1080/10635150490468675.

[58] J. Parker, A. Rambaut, and O. G. Pybus. Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infection, Genetics and Evolution*, 8(3):239–246, 2008. ISSN 1567-1348. doi: https://doi.org/10.1016/j.meegid.2007.08.001.

[59] W. Piel, M. Donoghue, and M. Sanderson. Treebase: a database of phylogenetic knowledge. *To the interoperable "Catalog of Life" with partners Species*, pages 41–47, 2000.

[60] S. Pompei, V. Loreto, and F. Tria. On the accuracy of language trees. *PloS one*, 6: e20109, 06 2011. doi: 10.1371/journal.pone.0020109.

[61] P. Ranacher, N. Neureiter, R. Gijn, B. Sonnenhauser, A. Escher, R. Weibel, P. Muysken, and B. Bickel. Contact-tracing in cultural evolution: a bayesian mixture model to detect geographic areas of language contact, 03 2021.

[62] M. Ravenhall, N. Škunca, F. Lassalle, and C. Dessimoz. Inferring horizontal gene transfer. *PLOS Computational Biology*, 11(5):1–16, 05 2015. doi: 10.1371/journal.pcbi.1004095.

[63] D. Robinson and L. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147, 1981. ISSN 0025-5564. doi: https://doi.org/10.1016/0025-5564(81)90043-2.

[64] A. Ronelle. Tracking sprachbung boundaries: Word order in the balkans. *Studies in Slavic and General Linguistics*, 28:9–27, 2000.

[65] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 07 1987. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a040454.

[66] A. Schleicher. *Die Darwinsche Theorie und Sprachwissenschaft: offenes Sendschreiben an Herrn Dr. Ernst Häckel*, volume 2. Böhlau, 1873.

[67] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[68] A. Stamatakis. Distributed and parallel algorithms and systems for inference of huge phylogenetic trees based on the maximum likelihood method, 2004.

[69] A. Stamatakis. Phylogenetics: Applications, software and challenges. *Cancer Genomics-Proteomics*, 2:301–305, 09 2005.

[70] M. Steel and D. Penny. Distributions of tree comparison metrics–some new results. *Systematic Biology - SYST BIOL*, 42:126–141, 06 1993. doi: 10.1093/sysbio/42.2.126.

[71] K. Strimmer and A. Rambaut. Inferring confidence sets of possibly misspecified gene trees. *Proceedings: Biological Sciences*, 269(1487):137–142, 2002.

[72] M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, 4(1), 06 2018. ISSN 2057-1577. doi: 10.1093/ve/vey016.

[73] M. Swadesh. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137, 1955.

[74] S. Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lect Math Life Sci (Am Math Soc)*, 17:57–86, 1986.

[75] A. Torres-Montúfar, T. Borsch, and H. Ochoterena. When homoplasy is not homoplasy: Dissecting trait evolution by contrasting composite and reductive coding. *Systematic Biology*, 67(3):543–551, 07 2017. ISSN 1063-5157. doi: 10.1093/sysbio/syx053.

[76] C. A. Villee. Morphology, 2023. URL `https://www.britannica.com/science/morphology-biology`.

[77] R. A. Vos, J. P. Balhoff, J. A. Caravas, M. T. Holder, H. Lapp, W. P. Maddison, P. E. Midford, A. Priyam, J. Sukumaran, X. Xia, and A. Stoltzfus. Nexml: Rich, extensible, and verifiable representation of comparative data and metadata. *Systematic Biology*, 61(4):675–689, 02 2012. ISSN 1063-5157. doi: 10.1093/sysbio/sys025.

[78] S. Wichmann, E. W. Holman, and C. H. Brown. The asjp database, 2022.

[79] Z. Yang. A space-time process model for the evolution of dna sequences. *Genetics*, 139(2):993–1005, 02 1995. ISSN 1943-2631. doi: 10.1093/genetics/139.2.993.

[80] Z. Yang. *Computational Molecular Evolution*. Oxford University Press, 10 2006. ISBN 9780198567028. doi: 10.1093/acprof:oso/9780198567028.001.0001.

[81] J. Zhang, G. A. Preising, M. Schumer, and J. A. Palacios. Crp-tree: A phylogenetic association test for binary traits, 2023.

Appendix

A. Geographical Trees

Here we provide cladograms for different geographical trees with original branch lengths. We obtain these trees with neighbor joining applied to the respective distance matrix.

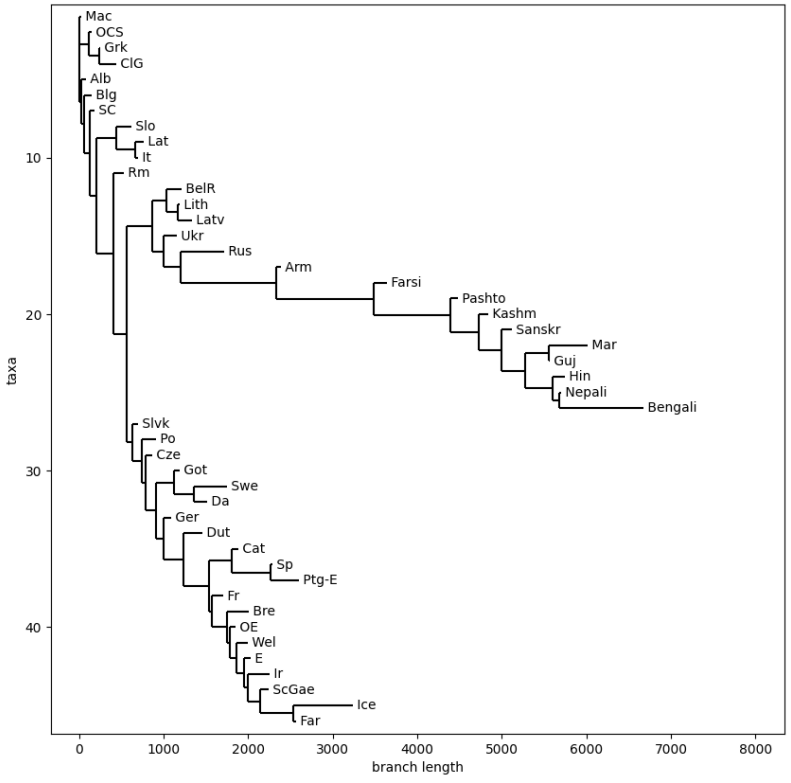


Figure A.1.: Haversine distances

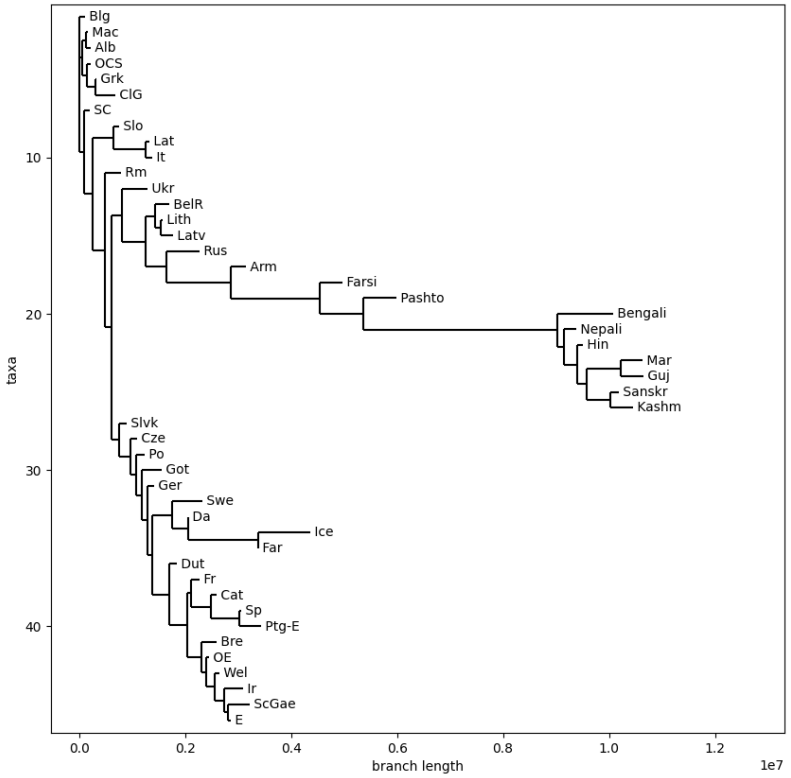


Figure A.2.: Route path lengths

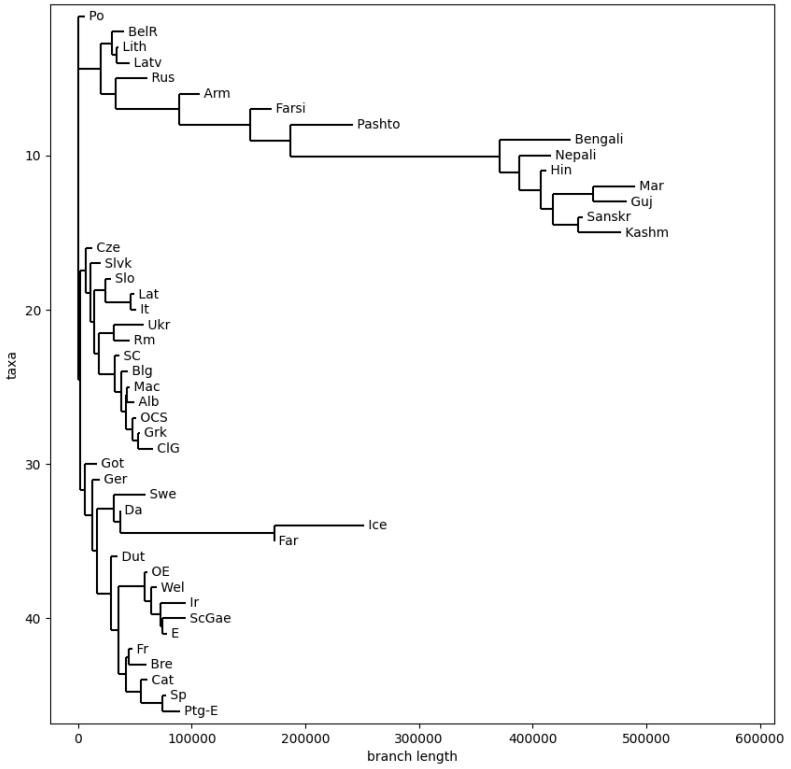


Figure A.3.: Route duration

B. Overview of Examined Linguistic Data Sets

Source	MSA Construction	Data Type	MSA Type	Number
Lexibank [48]	from CLDF	cognate	binary	70
			multi-value	14
		morphological	binary	4
			multi-value	0
SequenceComparison [50]	from CLDF	cognate	binary	8
			multi-value	3
phlorest [37]	as provided	cognate	binary	34
			multi-value	1
		morphological	binary	2
			multi-value	0
ASJP[42][78]	as provided	cognate	binary	65
	from clustering		multi-value	14
	as provided	sound class	binary	65
			multi-value	0
Tuled [35]	from CLDF	cognate	binary	1
			multi-value	0
ewave [43]	from CLDF	morphological	binary	1
			multi-value	1
sails [55]	from CLDF	morphological	binary	1
			multi-value	1
wals [20]	from CLDF	morphological	binary	0
			multi-value	0
DiACL [13]	from `.csv`	morphological	binary	3
			multi-value	0
Mouton Atlas[14]	from `.xlsx`	morphological	binary	1
			multi-value	0
bouckaert [10]	as provided	cognate	binary	1
			multi-value	0
wals (indo-europ.)[20, 53]	manually	morphological	binary	1
			multi-value	1
ouckaert(indo-europ.) [10, 53]	as provided	cognate	binary	1
			multi-value	0

B.1. Duplicates

Dataset in Lexibank	Dataset in phlorest	Remark
dyenindoeuropean	gray _and _atkinson2003	No Mapping for 3 taxa from phlorest and for 11 taxa from Lexibank
grollemundbantu	grollemund _et _al2015	
grollemundbantu	koile _et _al2022	No mapping for 5 taxa from Lexibank
kitchensemitic	kitchen _et _al2009	
powerma	power _et _al2020	
birchallchapacuran	birchall _et _al2016	
gerarditupi	gerardi _and _reichert2021	
utoaztecan	greenhill _et _al _subm	No mapping for 7 taxa from Lexibank
dravlex	kolipakam _et _al2018	
leekoreanic	lee2015	
leejaponic	lee _and _hasegawa2011	
leeainu	lee _and _hasegawa2013	
nagarajakhasian	nagaraja _et _al2013	
robinsonap	robinson _and _holton2012	
sagartst	sagart _et _al2019	No mapping for 2 taxa from phlorest and for 2 taxa from Lexibank
savelyevturkic	savelyev _and _robbeets2020	
mcelhanonhuon	greenhill2015	
bouckaert	bouckaert_et_al2012	The version bouckaert is from the supplementary of the paper, not from Lexibank

C. Software

The following enumeration lists the software and command lines we used for our analyses.

Chapter 2:

- RAxML-NG Version 1.1.0, available at `https://github.com/amkozlov/raxml-ng/releases/tag/1.1.0`, for retrieving intermediate trees and for the experiments with a new mixture model in Section 2.4, we used an adapted version based on this release, available at `https://github.com/luisevonderwiese/raxml-ng`
- standard-RAxML Version 8.2.12, available at `https://github.com/stamatak/standard-RAxML/releases/tag/v8.2.12`, used for the determination of the weight calibration in Section 2.2.2
- Root Digger Version v1.8.0-16-gc6d43e9, available at `https://github.com/computations/root_digger` in Section 2.2.3
- delta_statistics available at `https://github.com/mrborges23/delta_statistic` in Section 2.2.4
- IQ-TREE Version 2.2.0, available at `https://github.com/iqtree/iqtree2/releases/tag/v2.2.0`, used for the determination of plausible trees in Section 2.1.3 and Section 2.2.5 and for generating simulated MSAs with Alisim in Section 2.4

Chapter 3:

- Pythia training data pipeline as described in [39]. Note that we considered additional metrics compared to the publication that are implemented in a separate branch. The pipeline code is available at `https://github.com/tschuelia/difficulty-prediction-training-data/tree/tree_characterization`
- For running the above pipeline we used RAxML-NG Version 1.1.0, IQ-Tree Version 2.0.6, and Pythia Version 1.0.1.
- The database containing linguistic MSAs is available at `https://github.com/luisevonderwiese/language_alignment_database_interface`

Chapter 4:

- RAxML-NG Version 1.1.0, available at `https://github.com/amkozlov/raxml-ng/releases/tag/1.1.0`
- qdist available at `https://birc.au.dk/software/qdist` for the determination of generalized quartet distances

All scripts used for our analyses are available at `https://github.com/luisevonderwiese/mt_scripts`.