Are Sounds Sound for Phylogenetic Reconstruction?

Luise Häuser
Heidelberg Institute for Theoretical Studies
luise.haeuser@h-its.org

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

Taraka Rama
Independent Researcher
taraka.kasi@gmail.com

Johann-Mattis List
MPI-EVA / Univ. of Passau
mattis.list@uni-passau.de

Alexandros Stamatakis
Institute of Computer Science
FORTH
stamatak@ics.forth.gr

Abstract

In traditional studies on language evolution, scholars often emphasize the importance of sound laws and sound correspondences for phylogenetic inference of language family trees. However, to date, computational approaches have typically not taken this potential into account. Most computational studies still rely on lexical cognates as major data source for phylogenetic reconstruction in linguistics, although there do exist a few studies in which authors praise the benefits of comparing words at the level of sound sequences. Building on (a) ten diverse datasets from different language families, and (b) state-of-the-art methods for automated cognate and sound correspondence detection, we test, for the first time, the performance of sound-based versus cognate-based approaches to phylogenetic reconstruction. Our results show that phylogenies reconstructed from lexical cognates are topologically closer, by approximately one third with respect to the generalized quartet distance on average, to the gold standard phylogenies than phylogenies reconstructed from sound correspondences.

1 Introduction

Although controversially discussed in the beginning (Holm, 2007), quantitative approaches to phylogenetic reconstruction based on Bayesian phylogenetic inference frameworks have now become broadly accepted and used in the field of comparative linguistics. This is reflected by the increasing number of computer-based phylogenies that have been proposed for the world's largest language families – Dravidian (Kolipakam et al., 2018), Sino-Tibetan (Sagart et al., 2019), and Indo-European (Heggarty et al., 2023) – and even fully automated workflows, in which even cognate words are identified automatically, have shown to be comparatively robust (Rama et al., 2018). While rarely practiced in the pre-computational past of historical linguistics, computing detailed, fully resolved phylogenies with branch lengths and at times even estimated divergence times, has now become a routine tasks in contemporary language evolution studies.

Although traditional scholars have started to accept computational language phylogenies as a new tool deserving its place in the large tool chain of comparative linguistics, scholars still express substantial skepticism against most language phylogenies that have been inferred so far. One of the major arguments typically mentioned in this context is that phylogenetic approaches are usually based on cognate sets (sets of historically related words) that are identified in semantically aligned word lists. Since these *cognate sets* reflect *lexical data* only, many scholars mistrust them, given that lexical data are assumed to be substantially less stable over time than other aspects of languages (Campbell and Poser, 2008). Yet, for being able to infer stable phylogenetic trees a mix of conserved characters and more variable characters might be more beneficial.

In classical historical linguistics, the data used for subgrouping are traditionally composed of small collections of so-called *shared innovations* (Dyen, 1953). What counts as a shared innovation has itself never been clearly defined in the literature, but the largest amount of data used by scholars is traditionally taken from sound correspondences or supposed sound change processes (compare, for example the data in Anttila 1972, 305). Although it is controversially debated in the field (Ringe et al., 2002; Dybo and Starostin, 2008), many classical linguists still emphasize that sound correspondences are largely superior to lexical data to determine subgrouping.

There have only been few attempts to assess how well quantitative approaches to phylogenetic reconstruction perform when using sound correspondences instead of lexical cognates (Chacon and List, 2015). The main reason is that encoding

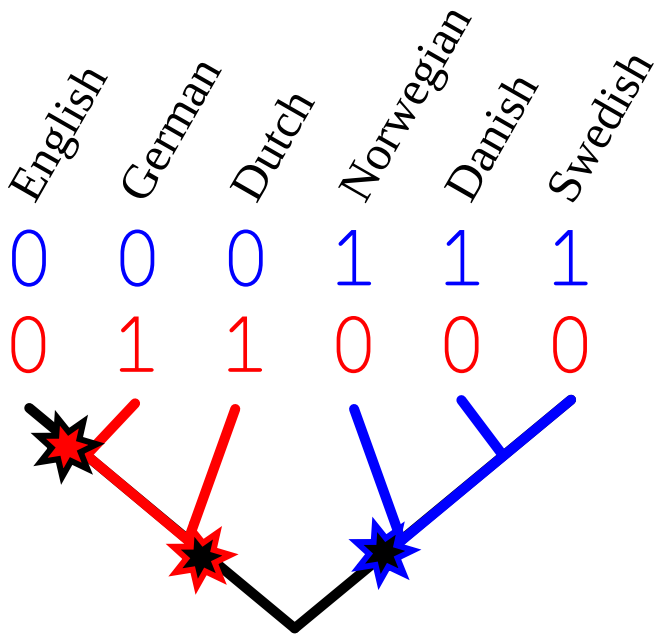
Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024), pages 78–87

Language	Concept	Form	Cog-Set
English	"big"	big	1
German	"big"	groß	2
Dutch	"big"	groot	2
Norwegian	"big"	stor	3
Danish	"big"	stor	3
Swedish	"big"	stor	3

(A) multi-state matrix

Concept		"big"		
Cog-Set		1	2	3
English	big	0	0	0
German	groß	0	1	0
Dutch	groot	0	1	0
Norweg.	stor	0	0	1
Danish	stor	0	0	1
Swedish	stor	0	0	1

(B) binary-state matrix



(C) evolutionary scenario (binary-state)

Figure 1: Gain-loss processes derived from binary cognate vectors. A shows a wordlist where cognate words are encoded as multi-state characters. B shows the corresponding binary encoding. C shows how gain and loss processes are modeled on a phylogenetic tree.

data to compute phylogenies from sound change patterns is tedious and labour-intensive even for a dataset comprising only 20 languages. Therefore, there have been but a few attempts to assess how well quantitative approaches to phylogenetic reconstruction perform when using sound correspondences instead of lexical cognates.

Here we build on state-of-the-art methods for automatic cognate detection and phonetic alignment in historical linguistics (List et al., 2016) and combine them with novel approaches for inferring sound correspondence patterns in multilingual datasets (List, 2019). Using this machinery we have devised a new workflow for phylogenetic reconstruction based on sound correspondence patterns. With a new collection of ten gold standard datasets, we test our workflow and compare it with alternative workflows that are exclusively based on lexical data. Our results indicate that sound correspondence patterns are substantially less suitable for the purpose of computer-based phylogenetic reconstruction than postulated.

2 Background

The majority of previous work on phylogenetic reconstruction using Bayesian phylogenetic inference (Kolipakam et al., 2018; Sagart et al., 2019; Rama et al., 2018) is based on cognate sets that are encoded as binary vectors. The presence or absence of a language in a cognate set is thus encoded as **1** or **0**, respectively. Subsequently, phylogenetic trees are inferred by assuming that cognate sets evolve along a phylogenetic tree via a gain and loss processes (see Figure 1).

The binary-state encoding is the most frequently used encoding technique; we deploy it in this study as well. Once such a dataset has been assembled, binary state data evolution can be modeled via a time-reversible binary state Continuous Time Markov Chain model (*binary-CTMC*, Bouckaert et al. 2012), which allows for gain and loss events to occur for an arbitrary number of times. Branch lengths on these trees reflect the mean number of expected substitutions (gain/loss events) per binary character site.

The *major contributions* of this study are: (1) We provide an automated workflow that allows to infer cognates and correspondence patterns and analyze them with the help of Bayesian phylogenetic inference methods, (2) we cross-validate the Bayesian inference results via Maximum Likelihood (ML) tree reconstructions and thereby discover that de-

Dataset	Words	Concepts	Languages	Distances	Sounds	Word Length
ConstenlaChibchan	1214	106	24	0.1	21.71	3.86
CrossAndean	2637	150	19	0.03	28.89	4.32
Dravlex	1341	100	20	0.06	36.85	4.53
FelekeSemitic	2412	150	19	0.05	45.32	4.99
HattoriJaponic	1710	197	10	0.03	34.9	4.47
HouChinese	1816	139	15	0.05	43	6.21
LeeKoreanic	1960	205	14	0.01	36.93	4.31
RobinsonAP	1424	216	13	0.03	24.38	4.51
WalworthPolynesian	6113	207	31	0.05	21.03	4.51
ZhivlovObugrian	1879	110	20	0.04	32.65	3.65

Table 1: Datasets and general aspects of the data. Distances refer to the average pairwise distance between all language pairs in the sample, derived from shared cognate counts (using the LingPy software). Number of sounds refers to the number of distinct sounds per language (on average), and the word length refers to the average length of the words observed in each dataset.

fault Bayesian priors that typically work well on molecular data can induce a prior bias when analyzing language datasets, (3) we show how the quality of phylogenetic reconstruction approaches based on sound correspondences can be compared to phylogenetic reconstruction based on lexical data, and in this way, and (4) we put the debate about the usefulness of sound-based as opposed to cognate-based phylogenies to the test.

As an early example for sound-based approaches to phylogenetic reconstruction, Hruschka et al. (2015) apply a CTMC model that allows for transitions between a fixed number of sounds for detecting the important sound changes in a dataset comprising etymologies across Turkic languages. Hruschka et al. do not infer phylogenies from their data. Instead, they use an established phylogeny (such established phylogenies are not readily available for many language families of the world) to infer branch lengths and transition probabilities between sounds in their data in order to detect sound changes at different time points in a time-calibrated family tree of Turkic.

Wheeler and Whiteley (2015) start from typical word lists (that would otherwise be used in phylogenetic reconstruction based on lexical data) and apply a parsimony-based algorithm that aligns words regardless if they are cognate or not, reconstructs a hypothetical ancestral word from the alignment, and seeks to infer the phylogeny that explains the observed sequences via the minimum amount of changes/mutations (Sankoff, 1975). In a later study, Whiteley et al. (2019) apply the same approach to a dataset of Bantu languages. The method by Wheeler and Whiteley (2015) is linguistically debatable, since words are not assigned to cognate sets prior to aligning them. It is well known that there is a strict difference between regular sound change processes and processes resulting from lexical replacement (Hall and Klein, 2010) and that even words that are cognate are not necessarily fully *alignable* (Schweikhard and List, 2020, 10).

Chacon and List (2015) start from manually extracted sound correspondence patterns for consonants in a dataset of 21 Tukanoan languages, to which proto-forms had also been manually added. Based on these sound correspondence patterns, they apply—in analogy to Wheeler and Whiteley (2015)—an algorithm that searches for the tree that provides the most parsimonious explanation for sound evolution. In contrast to Wheeler and Whiteley (2015), however, they added specific constraints for the transitions from one sound to another sound, which were based on expert judgments for the Tukanoan language family. The approach by Chacon and List (2015), finally, requires an enormous amount of preprocessing that entails the risk of inducing circular results, since proto-forms and major directions of sound change processes are required to be known in advance. While all approaches exhibit individual shortcomings, one of the largest shortcomings lies in the fact that it is very difficult to apply them systematically. This is also supported by the observation that no additional analogous studies have been conducted by other teams, despite the fact that all of the above methods have been proposed years ago.

3 Materials and Methods

3.1 Materials

In order to test whether sound correspondence patterns improve phylogenetic reconstruction or not, we selected ten datasets from the Lexibank repository (https://lexibank.clld.org, List et al. 2022) which were previously used to investigate the regularity of correspondence patterns in comparative cognate-coded wordlists (Blum and List, 2023). Lexibank offers published datasets in standardized formats (so-called Cross-Linguistic Data Formats, see Forkel et al. 2018). According to these standards, languages are linked to the Glottolog reference catalog (offering access to expert phylogenies and geolocations, https://glottolog.org, Hammarström et al. 2023), concepts are linked to the Concepticon reference catalog (offering fundamental definitions of semantic glosses and further information on concept properties, https://concepticon.clld.org, Concepticon), and sounds are provided in the phonetic transcription underlying the Cross-Linguistic Transcription Systems initiative (a reference catalog on speech sounds, offering a dynamic system that defines transcriptions for more than 8000 standard speech sounds observed in linguistic datasets, https://clts.clld.org, List et al. 2021; Anderson et al. 2018).

Data were preprocessed by first computing the phonetic alignment of all cognate sets in the data using the multiple alignment method proposed by List (2014). In a second step, these alignments were automatically *trimmed*, using the method proposed by Blum and List (2023), which identifies alignment columns with many gaps and ignores them, assuming that these result from morphological variation that would confuse cognate judgments.For phylogenetic inferences on molecular sequence data Tan et al. (2015) suggest that filtering worsens phylogenetic inference accuracy. The study by Blum and List (2023), however, shows that – for linguistic data – the overall regularity among cognates increases substantially, when trimming alignments systematically. Since regular sound correspondences provide the basis for the identification of classical sound laws that linguists typically use for the traditional subgrouping by shared innovations, we therefore consider the use of trimmed data as advantageous over using untrimmed alignments. Using trimmed phylogenies also has the advantage of reducing the noise, as can be seen from a rather drastic drop in the number of divergent sites in phylogenetic datasets that have been trimmed. However, it is beyond doubt that a closer investigation of the effects of trimming should be carried out in follow-up studies. In a third step, the method by List (2019) was used to compute correspondence patterns of the data. Phonetic alignments were conducted with LingPy (2.6.11, List and Forkel 2023a, https://pypi.org/project/lingpy). Trimming and correspondence pattern detection were carried out with LingRex (1.4.1, List and Forkel 2023b, https://pypi.org/project/lingrex).

Having identified correspondence patterns from the data, both the information on cognate sets and the information on correspondence patterns were converted into binary presence-absence matrices in Nexus format (Maddison et al., 1997), suitable for subsequent phylogenetic analysis.

3.2 Methods

Different methods for phylogenetic reconstruction have been described in the literature and have been controversially discussed among scholars for some time. Here we test two very basic approaches, Bayesian Inference and Maximum Likelihood. Since the data that we use for the inference of phylogenies comes in two flavors, derived as binary presence-absence matrices from cognate sets and from sound correspondence patterns, we test the methods on three different *character matrices*, namely the *cognate matrix*, derived from cognate judgments, the *sound correspondence matrix*, derived from sound correspondence patterns, and a *combined matrix*, in which we combine (concatenate) the cognate and the character matrix within a single new matrix.

In our experiments, we test three basic hypotheses. The first hypothesis assumes that phylogenetic inference on cognate sets is more accurate than phylogenetic inference based on sound correspondence patterns. The second hypothesis assumes that phylogenetic inference based on sound correspondence patterns is more accurate than phylogenetic inference based on cognate sets. The third hypothesis assumes that both character types do not differ substantially regarding their phylogenetic signal.

3.2.1 Bayesian Inference

Phylogenetic inferences were conducted using *MrBayes* (Ronquist and Huelsenbeck, 2003), version 3.2.7. For the final inferences presented here we

used the following prior settings for all datasets: (1) $\mathrm{Dirichlet}(1.0, 1.0)$ prior for base frequencies, (2) gamma-distributed rates, approximated by 4 discrete categories, with standard exponential prior for the shape of the gamma distribution that models among site rate heterogeneity, (3) uniform prior over tree topologies, and (4) strict clock model of branch lengths.

We initially used an an exponential distribution with a rate of 1.0 as prior for the Γ model of rate heterogeneity. This prior constrains the α shape parameter of the Γ distribution to relatively small values. As a consequence, MrBayes obtains α values below 10 for almost all data sets and posterior samples it draws. This indicates a high to moderate degree of rate heterogeneity. However, our ML analyses (see below) yielded substantially higher ML estimates for α on some datasets. To investigate this discrepancy, we repeated the Bayesian inferences, now using $\mathrm{Uniform}(0.01, 100)$ as a prior for the Γ distribution of rate heterogeneity. As a consequence, we obtained a different distribution of the α values that better reflects the corresponding ML estimates.

The more informative default exponential prior in MrBayes has presumably been developed for molecular datasets, which usually exhibit a high degree of rate heterogeneity. In other words, ML estimates of α exhibit a small variance (see, e.g., https://github.com/angtft/RAxMLGroveScripts/blob/main/figures/test_ALPHA.png and the corresponding paper by Höhler et al. (2021)). When executing inferences on language datasets, using this default molecular prior can hence bias the results. That is, had we not conducted complimentary ML analyses, this surprising dataset-dependent bi-modal distribution of α values on language datasets (see Table 2) would have gone unnoticed. We thus strongly advocate that all default priors for molecular datasets should be carefully and critically re-assessed when conducting Bayesian inferences on language datasets and that ML analyses should always complement Bayesian Inferences.

Motivated by this observation, the Bayesian analysis was repeated, now using a uniform prior over the interval $[0.01, 100.0]$ for α.

We sampled the state of the Markov chain every 1,000th generation. We stopped MCMC chains when the average standard deviation of split frequencies (ASDSF) was below 0.01 after discarding the first 25% of the samples.[1]

The median posterior value for α are shown in Table 3. From the remaining 75% of the recorded samples from the two cold chains, 1,000 trees were drawn at random and used for further evaluation.

If one of the two individual character types provides the best results, this would be evidence for Hypothesis 1 or Hypothesis 2. If the combined dataset provides the best results, this would be evidence for Hypothesis 3.

To evaluate the quality of the inferred phylogenies, we used the classifications from Glottolog (Hammarström et al., 2023). The topological distance or degree of consistency of an inferred strictly binary (fully bifurcating) phylogeny and a (potentially polytomous/multi-furcating) Glottolog tree was measured as the *generalized quartet distance* (GQD), as proposed in (Pompei et al., 2011).[2]

3.2.2 Maximum Likelihood Tree Inferences

To exclude any potential bias by the selected tree inference method, we also conducted independent Maximum Likelihood (ML) tree inferences. For ML tree inference we used RAxML-NG (Kozlov et al., 2019), version 1.2.0. For each dataset and character matrix type (cognate/sound/concatenated) we executed 20 independent ML tree searches using the default tree search configuration of RAxML-NG (10 searches starting from random trees and 10 searches starting from randomized stepwise addition order parsimony trees) under the BIN+G model of binary character substitution with ML estimated base frequencies. We approximate the Γ model of rate heterogeneity via four discrete rates. Thus, each inference includes the ML estimate of the $\alpha \in [0.0201, 100]$ shape parameter that determines the shape of the Γ distribution. The smaller the estimate of α, the higher the rate heterogeneity in the respective dataset will be (Yang, 1995). For three matrices containing cognate data and for two matrices encoding sound correspondences, we

[1] Note that convergence diagnosis metrics such as ASDSF can only serve to diagnose the failure of an MCMC chain to converge, but can never confirm its convergence.

[2] The GQD is a generalization of the well-known *quartet distance* (Estabrook et al., 1985) that allows to compare fully bifurcating trees with multi-furcating trees. The GQD is defined as the number of quartets that are not shared between the two trees, divided by the number of all possible quartets. The GQD is a number between 0 and 1, where 0 means that the two trees are identical, and 1 means that the two trees are completely different.

Dataset	Cognates	Sound Correspondences	Combined
ConstenlaChibchan	0.592	**99.871**	4.178
CrossAndean	1.243	6.334	1.154
Dravlex	0.702	4.301	2.234
FelekeSemitic	1.062	7.430	2.693
Hattorijaponic	**99.848**	**99.897**	**99.890**
HouChinese	2.357	6.120	4.195
LeeKoreanic	8.316	8.420	3.284
RobinsonAP	**99.869**	15.269	3.486
WalworthPolynesian	1.333	4.233	1.624
ZhivlovObugrian	**99.850**	4.244	3.134

Table 2: ML estimates of the alpha shape value of the Gamma model for among site rate heterogeneity for all languages and all character matrices. Values indicating an extremely low rate heterogeneity (all sites evolve at the same rate) are highlighted in bold.

Dataset	Cognates	Sound Correspondences	Combined
ConstenlaChibchan	1.758	**53.115**	1.138
CrossAndean	1.620	19.558	0.400
Dravlex	0.749	23.613	0.814
FelekeSemitic	0.932	41.669	0.727
HattoriJaponic	**58.012**	**60.602**	0.268
HouChinese	3.011	27.476	0.933
LeeKoreanic	**52.045**	39.354	0.058
RobinsonAP	**56.928**	**51.818**	0.373
WalworthPolynesian	1.480	4.348	0.800
ZhivlovObugrian	**58.652**	**51.280**	0.507

Table 3: Median Bayesian estimates of the alpha shape value of the Gamma model for among site rate heterogeneity for all languages and all character matrices. Values indicating an extremely low rate heterogeneity (all sites evolve at the same rate) are highlighted in bold.

obtain an estimate for $\alpha > 99.8$, which means that all sites evolve at the same rate and that there is essentially no rate heterogeneity. Hence, trees on these datasets could also be inferred without correcting for rate heterogeneity. For the remaining datasets, the ML estimates of α are below 20, indicating a moderate to high degree of rate heterogeneity. This extreme bi-modal distribution of α estimates differs substantially from the distribution we observe on tens of thousands of empirical (i.e., non-simulated) molecular datasets (Höhler et al., 2021) (also see the respective distribution of $alpha$ values plot at https://github.com/angtft/RAxMLGroveScripts/blob/main/figures/test_ALPHA.png where $alpha$ values range approximately between 0.01 and 1.5).

We have currently not been able to identify which intrinsic dataset properties cause this surprising and extreme bi-modal distribution of $alpha$ values in language datasets. For the given datasets, we examined the number of concepts and languages under study, the dimensions of the MSAs and the average branch lengths in the trees inferred. We determined the difficulty score using Pythia (Haag et al., 2022) and the number of species using the method for species delimitation implemented in mPTP (Kapli et al., 2017). For none of these properties we were able to find a clear connection to the value estimated for α. One path to explore in analogy to molecular biology is whether some language datasets should be regarded as representing individuals from a population of the same species while others represent distinct species. In fact, for molecular data we have thus far only observed such high estimates of α values (i.e., low or no rate heterogeneity) for population genetic datasets comprising

sequences of individuals of the same species or closely related sub-species.

3.3 Implementation

Methods for data handling and preprocessing are implemented in Python (with specific requirements and software packages indicated above), R and Julia. For the phylogenetic analyses, dedicated third party packages are used. All information on how to replicate our study and how to inspect individual analyses are provided in the supplementary material accompanying this study.

4 Results

4.1 Bayesian Inference

We computed the GQD to the goldstandard tree for each of the 1,000 samples from the posterior and computed the median for each dataset and character type. The results of our evaluation are shown in Table 4. As can be seen from the table, phylogenetic inferences based on cognate class data and on concatenated cognate/sound data provide results that are about equally good, with a slight advantage for concatenated data. Phylogenetic inference based on sound correspondences alone yields results that are clearly worse. The concatenated dataset provides the best results for seven out of ten datasets, while in three cases, the cognate class dataset provides the best results. The sound correspondence dataset never yields the best results. These results provide clear evidence in favor of Hypothesis 1 and against Hypothesis 2. The decision about Hypothesis 3 is somewhat equivocal.

4.2 Maximum Likelihood

Table 5 shows the evaluation results for the ML based inferences. Note that we obtain slightly different results when calculate the average distance to all 20 inferred ML trees or when using the BIN model (without accounting for among site rate heterogeneity). The corresponding GQ distances can differ by up to 0.17, although differences of > 0.05 only occur for 7 of the 30 MSAs under study. However, the following observations apply in all cases. First, there is no dataset where the tree inferred on the sound correspondences is substantially closer to the gold standard than the trees inferred on the cognate or concatenated data. On the other hand, there are three datasets (CrossAndean, HouChinese, LeeKoreanic) for which inferences on sound correspondence data yield trees with a substantially higher GQ distance to the gold standard. Inferences on the cognate and combined datasets yield comparable distances to the gold standard. Hence, the results of our ML analyses are consistent with the Bayesian inference results.

5 Discussion and Conclusion

While our results are less conclusive than one might expect, we think that they show clearly enough that sound-correspondence-based phylogenies should be taken with care. As we show, sound-correspondence-based phylogenies do rarely substantially outperform cognate-based phylogenies. Instead, we observe that cognate-based phylogenies are topologically much closer to the gold standard on average. At this point, we cannot say, whether combined approaches significantly outperform phylogenies purely inferred from cognate sets. Future studies that expand the data we used in this study are needed to clarify this question.

Given the prior bias we observed for default parameter priors that work well for Bayesian inference on molecular data, we advocate for a critical re-assessment of all priors that are being routinely used in Bayesian analyses of language data. This reassessment can be conducted by routinely executing analogous ML analyses and carefully inspecting all ML parameter estimates (branch lengths, tree length, base frequencies, α shape parameter) and not only focusing on the resulting tree topology. We also cross-checked the estimates for the base frequencies, but did not observe any discrepancies between ML and Bayesian Inference as a flat default ($\beta(1, 1)$) prior was used. The reasons for the extreme bi-modal distribution of α values we observed remain unclear, despite the fact that we have assessed 30 different dataset characteristics and summary statistics that are, however, all uncorrelated with the α estimate. Using machine learning techniques to predict α values for datasets and thereby potentially understand the dataset properties responsible for this bi-modal distribution is not feasible due to an insufficient amount of available data. Investigating this issue hence remains subject of future work.

Supplementary Material

The supplementary material including data and code necessary to replicate the experiments discussed in this study along with instructions on how to run the code are curated

Dataset	Cognates	Sound Correspondences	Concatenated
ConstenlaChibchan	0.245	0.414	**0.212**
CrossAndean	**0.148**	0.523	0.189
Dravlex	0.336	0.351	**0.320**
FelekeSemitic	**0.083**	0.146	0.113
HattoriJaponic	0.585	0.431	**0.362**
HouChinese	**0.240**	0.494	0.377
LeeKoreanic	0.224	0.358	**0.157**
RobinsonAP	0.424	0.281	**0.259**
WalworthPolynesian	0.179	0.252	**0.146**
ZhivlovObugrian	0.330	0.356	**0.316**
median	0.251	0.358	**0.240**

Table 4: Generalized quartet distances (posterior medians) for Bayesian inference. The best result for each dataset is highlighted in bold.

Dataset	Cognates	Sound Correspondences	Combined
ConstenlaChibchan	0.335	0.360	**0.283**
CrossAndean	0.246	0.470	**0.088**
Dravlex	0.358	0.472	**0.307**
FelekeSemitic	0.126	**0.103**	0.126
HattoriJaponic	**0.532**	0.681	0.559
HouChinese	0.224	0.529	**0.186**
LeeKoreanic	**0.178**	0.386	0.204
RobinsonAP	0.355	**0.321**	0.348
WalworthPolynesian	**0.139**	0.188	0.192
ZhivlovObugrian	**0.322**	0.356	0.360
median	0.284	0.373	**0.243**

Table 5: Generalized quartet distances between the gold standard trees and the the best-scoring ML tree inferred under the **BIN+G** model. The best result for each dataset is highlighted in bold.

on GitHub (https://github.com/lingpy/are-sounds-sound-paper) and archived with Zenodo (https://doi.org/10.5281/zenodo.10610428).

Limitations

The ongoing debate of what evidence phylogenetic reconstruction should be based on cannot be considered as conclusively solved with this study, although we are confident that our contribution merits the attention of all scholars participating in the debate. One crucial weakness of our approach, which we cannot overcome completely at the moment, is the way we operationalize "sound laws as evidence for phylogenetic reconstruction". Here, we use sound correspondence patterns which we infer automatically from the data sets. One may criticize that this procedure is not identical with the way in which experts do cladistic subgrouping. In response to such criticism, we emphasize, however, that every attempt to arrive at a useful way to compare evidence based on sound correspondence patterns (and sound laws) with evidence based on cognate sets, must start at some point, and that we are convinced that this approach comes quite close to the evidence traditional scholars defending phylogenetic reconstruction by innovation have in mind.

Acknowledgments

This research was supported by the Max Planck Society Research Grant *CALC*^{3} (JML, https://digling.org), the ERC Consolidator Grant *ProduSemy* (JML, Grant No. 101044282,

see https://doi.org/10.3030/101044282), the ERC Advanced Grant *CrossLingFerence* (GJ, Grant. No. 834050, see https://doi.org/10.3030/834050), the Klaus-Tschira Foundation, and by the European Union (EU) under Grant Agreement No 101087081 (AS, Comp-Biodiv-GR, see https://doi.org/10.3030/101087081). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them. We thank Maria Heitmeier and Harald Baayen for their valuable input regarding the computational models used in this study.

References

Cormac Anderson, Tiago Tresoldi, Thiago Costa Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A Cross-Linguistic Database of Phonetic Transcription Systems. *Yearbook of the Poznań Linguistic Meeting*, 4(1):21–53.

Raimo Anttila. 1972. *An introduction to historical and comparative linguistics*. Macmillan, New York.

Frederic Blum and Johann-Mattis List. 2023. Trimming phonetic alignments improves the inference of sound correspondence patterns from multilingual wordlists. In *Proceedings of the 5th Workshop on Computational Typology and Multilingual NLP*, pages 52–64. Association for Computational Linguistics.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard, and Quentin D Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

Lyle Campbell and William J. Poser. 2008. *Language Classification: History and Method*. Cambridge University Press.

Thiago Costa Chacon and Johann-Mattis List. 2015. Improved computational models of sound change shed light on the history of the Tukanoan languages. *Journal of Language Relationship*, 13(3):177–204.

Anna Dybo and George S Starostin. 2008. In defense of the comparative method, or the end of the Vovin controversy. In I. S. Smirnov, editor, *Aspekty komparativistiki*, pages 119–258. RGGU, Moscow.

Isidore Dyen. 1953. [Review] Malgache et maanjan: Une comparaison linguistique by Otto Chr. Dahl. *Language*, 29(4):577–590.

George F Estabrook, FR McMorris, and Christopher A Meacham. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology*, 34(2):193–200.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(180205):1–10.

Julia Haag, Dimitri Höhler, Ben Bettisworth, and Alexandros Stamatakis. 2022. From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses. *Molecular Biology and Evolution*, 39(12):msac254.

David Hall and Dan Klein. 2010. Finding cognate groups using phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1030–1039. Association for Computational Linguistics.

Harald Hammarström, Martin Haspelmath, Robert Forkel, and Sebastian Bank. 2023. *Glottolog. Version 4.8*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard F. Strand, Geoffrey Haig, Martin Macák, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, Matthew Boutilier, Cassandra Freiberg, Robert Tegethoff, Matilde Serangeli, Nikos Liosis, Krzysztof Stroński, Kim Schulte, Ganesh Kumar Gupta, Wolfgang Haak, Johannes Krause, Quentin D. Atkinson, Simon J. Greenhill, Denise Kühnert, and Russell D. Gray. 2023. Language trees with sampled ancestors support a hybrid model for the origin of indo-european languages. *Science*, 381(6656).

Hans J. Holm. 2007. The new arboretum of Indo-European trees. *Journal of Quantitative Linguistics*, 14(2-3):167–214.

Daniel J Hruschka, Simon Branford, Eric D Smith, Jon Wilkins, Andrew Meade, Mark Pagel, and Tanmoy Bhattacharya. 2015. Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology*, 25(1):1–9.

Dimitri Höhler, Wayne Pfeiffer, Vassilios Ioannidis, Heinz Stockinger, and Alexandros Stamatakis. 2021. RAxML Grove: an empirical phylogenetic tree database. *Bioinformatics*, 38(6):1741–1742.

P Kapli, S Lutteropp, J Zhang, K Kobert, P Pavlidis, A Stamatakis, and T Flouri. 2017. Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. *Bioinformatics*, 33(11):1630–1638.

Vishnupriya Kolipakam, Fiona M. Jordan, Michael Dunn, Simon J. Greenhill, Remco Bouckaert, Russell D. Gray, and Annemarie Verkerk. 2018. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science*, 5(171504):1–17.

Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21):4453–4455.

Johann-Mattis List. 2014. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

Johann-Mattis List. 2019. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics*, 1(45):137–161.

Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021. *Cross-Linguistic Transcription Systems. Version 2.1.0*. Max Planck Institute for the Science of Human History, Jena.

Johann-Mattis List and Robert Forkel. 2023a. *LingPy. A Python library for quantitative tasks in historical linguistics [Software Library, Version 2.6.9]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List and Robert Forkel. 2023b. *LingRex: Linguistic reconstruction with LingPy*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(316):1–31.

Johann-Mattis List, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*, pages 599–605, Berlin. Association of Computational Linguistics.

D. R. Maddison, D. L. Swofford, and W. P. Maddison. 1997. NEXUS: an extensible file format for systematic information. *Syst. Biol.*, 46(4):590–621.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS one*, 6(6):e20109.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400.

Donald Ringe, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Frederik Ronquist and John P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of sino-tibetan. *Proceedings of the National Academy of Science of the United States of America*, 116:10317–10322.

David Sankoff. 1975. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42.

Nathanael E. Schweikhard and Johann-Mattis List. 2020. Developing an annotation framework for word formation processes in comparative linguistics. *SKASE Journal of Theoretical Linguistics*, 17(1):2–26.

Ge Tan, Matthieu Muffato, Christian Ledergerber, Javier Herrero, Nick Goldman, Manuel Gil, and Christophe Dessimoz. 2015. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Systematic Biology*, 64(5):778–791.

W. C. Wheeler and Peter M. Whiteley. 2015. Historical linguistics as a sequence optimization problem: the evolution and biogeography of uto-aztecan languages. *Cladistics*, 31(2):113–125.

Peter M. Whiteley, Ming Xue, and Ward C. Wheeler. 2019. Revising the bantu tree. *Cladistics*, 35:329–348.

Z Yang. 1995. A space-time process model for the evolution of dna sequences. *Genetics*, 139(2):993–1005.